VARIABLE SELECTION AND PREDICTION FOR COMPLEX SURVIVAL DATA ANALYSIS

Xiaowei Ren

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

July, 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Shanshan Li, Ph.D., Co-Chair

Zhangsheng Yu, Ph.D., Co-Chair

Doctoral Committee

Wanzhu Tu, Ph.D.

May 17, 2017

Gregory K. Steele, Ph.D.

DEDICATION

To My Beloved Family.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my advisors Dr. Zhangsheng Yu and Dr. Shanshan Li for their constant guidance, encouragement and support in my Ph.D. study. I really appreciate the opportunity that they lead me to grow in this promising research area. Their instruction has not only trained my knowledge and expertise, but also cultivated me with open-mindedness, ability of critical thinking, and skills of effective communication in both oral and written that are essentially critical for my future success. I have also learned from them the spirit of hard working, persistence, patience and creativity, which I will benefit from for the rest of my life.

I would like to thank the committee members for my thesis research, Dr. Wanzhu Tu and Dr. George Steele for their critical evaluations on my dissertation. I would also like to specially thank Dr. Giorgos Bakonyannis for his particular personal advice in competing risks area where his expertise can effectively help me keep my research in correct direction. Very special thanks to Dr. Ying Zhang for his short but helpful advice of the results presentation of my second topic.

Finally, I want to thank all my family and friends for their encouragement throughout the years.

Xiaowei Ren

VARIABLE SELECTION AND PREDICTION FOR COMPLEX SURVIVAL DATA ANALYSIS

Survival analysis methods for time-to-event data are commonly used in biomedical researches. It is essential to select the important variables and identify the correct covariate functional form. After selection of important variables, it is of interest to evaluate the prediction performance of the selected model, typically by receiver operating characteristic (ROC) curve. Furthermore, the analysis of time-to-event data is complicated by the presence of interval censoring and dependent competing events, both of which occur frequently in clinical studies. In this dissertation, we set to develop variable selection and prediction methods for complex survival data. In the first topic, we proposed a two-stage procedure to identify the linear and/or non-linear covariates functional forms simultaneously and estimate the selected covariate effects for competing risks data. Spectral decomposition was used to decompose the nonparametric covariate function. The adaptive LASSO method was then to select the linear and non-linear components, respectively. We showed that our method achieved good selection accuracy and minimal estimation biases. In the second topic, to evaluate the prediction performance, we extended the ROC function estimation of right-censored competing risks data to interval-censored data. We proved the consistency of the estimator and demonstrated the convergence of estimator in numerical studies. In the third topic, we extended the ROC function for independent survival data to clus-

tered survival data using within-cluster-resampling (WCR) technique. All the three methods had been implemented in real data as illustration.

Shanshan Li, Ph.D., Co-chair

Zhangsheng Yu, Ph.D., Co-chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Survival analysis methods are widely used to analyze the time from entry to occurrence time of events of interest. The techniques developed in survival analysis are now applied in many fields, such as biology, engineering, medicine, quality control, credit risk modeling in finance. On the other hand, statistical modeling plays an increasingly important role in modern scientific investigation. An important problem in survival data modeling is how to model the conditional hazard rate of failure times given certain covariates via a regression model. These problems have presented a significant challenge to statisticians in the last five decades. The validity of model-based scientific inquiry is usually contingent on the correct specification of the model. Failure to include relevant independent variables will result in questionable inference, while including irrelevant variables creates numerical instability and reduces analytical efficiency. Determination of the correct model structure based on observed data has therefore become an essential component of the modeling process. Ultimately, one hopes to achieve a parsimonious modeling structure without sacrificing predictive or explanatory power. After determining the best model using appropriate variable selection method, people often need to evaluate the effectiveness of the selected model in terms of prediction accuracy. With a large number of potential predictors, variable selection is often the first step in developing predictive models. There are many reasons for focusing on a subset: the desire to glean important biological insight,

operational considerations for how this information can be utilized in subsequent development of the novel therapy, and the fact that a simple model has a better chance to hold in a new trial and lends itself more easily to validation efforts. Suppose that we have a clinical response variable Y, and a set of p predictors $X_1, \ldots, X_p$. The problem of variable selection arises when we want to model the relationship between Y and a subset of $X_1, \ldots, X_p$. Variable selection methods have mainly been developed for linear models. Heuristic variable selection procedures are often employed, for example, forward selection, backward elimination, and stepwise selection. A number of well-known selection criteria have been used in heuristic procedures for linear models, including AIC, BIC, and Mallows' Cp. More recently, regularization methods have also been used as variable selection approaches, for example, LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), and elastic net (Zuo and Hastie, 2005). The regularization methods conduct variable selection mainly in linear or parametric models. However, the linear relationship between response variable and one or more predictors is often too simple to be proper in complicated clinical data analysis. The limitations of the linear model and other parametric statistical approaches motivate our use of nonparametric methods to model the relationship between Y and $X_1, \ldots, X_p$.

The objective of this dissertation is to develop a procedure of linear and non-linear variable selection and evaluation procedure for proportional hazards models of survival data in complex settings. In this chapter, I present my research questions, review the existing literature, and describe the general approaches that I use in this research.

1.1 Cumulative Incidence Function Modeling in Competing Risks Data

The Cox proportional hazards model and its associated partial likelihood estimation method (Cox, 1972) has stimulated a lot of works in this field. Cox proportional hazards model has been extended to various settings of survival data analysis, including competing risks data, clustered survival data, recurrent events data, interval censored data, etc. A key quantity in the analysis of competing risks data is the cause-specific cumulative incidence function (CIF). The cause-specific CIF gives the probability of experiencing a particular event as a function of follow-up time, accounting for the fact that some individuals will not have the event of interest because of experiencing a competing event. The cause-specific CIF is a measure of absolute risk and the estimation of cause-specific CIF does not require independence assumption of the competing events. Only if the competing risks are independent is the estimator of the cause-specific hazard equal to the estimator of the marginal hazard. Therefore, Fine and Gray (Fine and Gray, 1999) proposed proportional sub-distribution hazards model that links the covariate effect and cause-specific CIF and facilitates the direct modeling of marginal hazard of the event via the proposed "sub-distribution hazard".

There are two main approaches to statistical modelling of competing risks data: (i) modelling cause-specific hazards and (ii) direct modelling of the cause-specific CIF. In the first approach, both the cause-specific hazard function of the primary event of interest and any competing events must be modelled if an estimate of the cause-specific CIF is required. Thus, with one event of primary interest and one competing event, two cause-specific hazard models are required. For the second approach of

direct modelling of the cause-specific CIF, the most common method is to use the proportional sub-distribution hazards model first described by Fine and Gray. The model is semi-parametric in that the baseline sub-distribution hazard function is not directly estimated with the only parameters directly estimated being log sub-distribution hazard ratios. Jeong and Fine developed a parametric approach where cause-specific CIFs could be directly estimated by simultaneously fitting models to both causes using a Gompertz distribution with a generalized link function that has proportional sub-distribution hazards models and proportional odds models as special cases. An extension of this work proposes a parameterization that constrains the sum of the cause-specific CIFs to not exceed 1 (Shi et al., 2013). For example, in the illustration presented by Choi and Huang (Choi and Huang, 2014), the analysis based on the Fine-Gray model produced cumulative incidence estimates that added to 1.29. This may indicate a model misspecification and may adversely influence the validity of any predictions based on these estimates.

1.2 Linear and Non-linear Variable Selection in Competing Risks Data

The proportional hazards models assuming linear covariate effects are widely used for regression in survival analysis. However, the linear assumption of covariates effect may be too rigid and unrealistic to depict the covariates effect on hazard. As a result, the proportional hazards models involving splines or local kernel method have been studied by many authors (Fan et al., 1997; Yu and Lin, 2008) in that basis expansion and regression spline methods are popular nonparametric techniques used to characterize nonlinear covariate effects (Gray, 1992; O'Sullivan, 1988, 1993). The Cox

model has extended the parametrization of covariates from linear to non-linear using non-parametric modeling technique. One big challenge in survival analysis applied in clinical research is how to efficiently select the important variables. Furthermore, in clinical research it is common to have non-linear effect of covariate in modeling, which may not be identified if the covariate effect is assumed to be linear. For example, the log-relative hazard function of age has a convex shape in a Veteran's Administration lung cancer trial (Kalbfleisch and Prentice, 2002). To estimate nonlinear covariate effects, one popular approach is to characterize the nonlinear effects using basis expansion and regression spline methods (O'Sullivan, 1988, 1993). However, when the underlying covariate effect is linear or zero (no effect), using a regression spline often complicates the interpretation of analysis results. Therefore, investigators often face a dilemma between balancing model complexity and model goodness-of-fit when using nonparametric covariate functions. The more basis functions are used, the better one can approximate the true underlying function, but it is at the price of increasing both computation intensity and model complexity. It is of interest to develop a data-driven approach to select the covariate functional form(or linear and non-linear structure) for competing risk model using variable selection approaches. This sort of issue requires data-driven method for selecting the linear and nonlinear effects. In this topic, we adopt spectral decomposition of Wand and Ormerod (2008) and adaptive LASSO of Zou (2006) to perform the variable selection and structure discovery in competing risks setting. Our method not only identifies the true effect of variables, but also bridges the connection between variable selection and non-parametric estimation of Fine-Gray sub-distribution hazard model. This connection provides more flexibility for variable screening and variable effects estimation. We apply the two

stage selection and estimation procedure (Zhang et al., 2011) to simultaneously select the covariate coefficents.

1.3 Estimation of Time-dependent ROC Curves with Interval Censored Survival Data in Competing Risks Setting

A logical question following variable selection and structure discovery is how well this selected variables can predict the outcome. After selection of important variables, it is natural for investigators to evaluate the prediction accuracy of the selected model. In survival data analysis, since the outcome is the predicted survival function, the ultimate outcome in terms of prediction is the failure/non-failure, which can be further interpreted as case/control if we consider a certain event's occurrence as case. Because censored data share feature of both continuous response data and binary data, the accuracy concepts that are standard for either response type are extended to application on survival outcomes. For continuous predictor variable, there are typically two common approaches for assessment of predictors, the proportion of variation explained by the covariates (R^2), and receiver operating characteristics (ROC) function. Recently ROC application has been extended to survival analysis including the nonparametric approach by Heagerty et al. (2000) and the semi-parametric approach by Heagerty and Zheng (2005). Previous research has focused on extending the proportion of variation explained by the covariates, or R^2, to censored data models (O'Quigley and Xu, 2001; Schemper and Henderson, 2000). Time-dependent ROC curves offer an alternative to the use of R^2 extensions for survival data. However, the goal of an ROC analysis is to characterize the prognostic potential of a marker (or model) by

focusing on the correct classification rates. Methods that summarize the proportion of variation explained by covariates require a different estimation approach, and have a different ultimate objective. Therefore, people investigated the ROC function applied on survival modeling with various settings, such as competing risks (Saha and Heagerty, 2010; Zheng et al., 2012). Careful literature search does not yield published work in ROC function estimation in interval censored competing risks data settings. Therefore, in this topic, I plan to extend the ROC function estimation method for right censored data to the interval censored survival data.

1.4 Estimation of Time-dependent ROC Curves with Clustered Survival Data

The ROC curve for survival data has been extensively studied, including semiparametric approach based on two different ROC function definitions using proportional hazard model. However, to the best of our knowledge, ROC curve estimation for the clustered survival data is still an open question. The diagnostic studies in which each patient has several diseased and nondiseased observations generate clustered ROC data. Within the same cluster, observations are naturally correlated, and the cluster size may be random. The traditional ROC methods on clustered data can result in a biased variance estimator and subsequently lead to incorrect statistical inference. We introduce resampling methods on clustered ROC data to account for the within-cluster correlation. The within-cluster resampling ROC methods work as follows. First, one observation is randomly selected from each patient/cluster, and then the traditional ROC methods are applied on the resampled data to obtain resampled ROC estimates. These steps are performed many times and the average of resampled

ROC estimates is the final estimator. The proposed methods do not require a specific within-cluster correlation structure and yield valid estimators while accounting for the within-cluster correlation. We compare the standard error estimated using our approach and naive approach which treats clustered data as independent. Also, for the within-cluster resampling estimate, there is established large sample property (Hoffman et al., 2001). Given the well built asymptotic normality and the straightforward operation procedure of the within-cluster resampling technique, in the third topic, I focus on applying the within-cluster-resampling technique to facilitate the estimation of ROC function in clustered survival data. Numerical studies demonstrate that the proposed estimation performs well with practical sample sizes. Application to the diabetic retinopathy study data (Sorbinil Retinopathy Trial Research Group, 1990) is given as an illustration.

From the three chapters above, we can see the wide application potential of the survival technique in competing risks data settings. How to systematically develop a procedure to implement the variable selection and evaluation procedure in complex survival data setting is of imperative importance and meaningfulness in clinical research plan and read data analysis. I would like to study and practice the development of this procedure using the above three topics to illustrate the feasibility of such implementation. This dissertation addresses this need in a systematic way, by proposing a integrated procedure that shows an example to address these demand.

Chapter 2

Linear and Non-linear Variable Selection in Competing Risks Data

2.1 Introduction

Competing risks data arise when study subjects are at risk of more than one types of events of interest, and the occurrence of one event may prevent the occurrence of other potential events, thus only the earliest event can be observed. In competing risks analysis, covariates often affect different events in different ways. For example, the Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) Study (The AFFIRM Investigators, 2002, 2004) compared the overall mortality between rhythm-control and rate-control treatments for patients with atrial fibrillation. In this study, subjects may experience death caused by cardiovascular disease (e.g. myocardial infarction), or other fatal non-cardiovascular complications (e.g. cardiogenic shock) which can be viewed as competing events for cardiovascular death. The observed event "death from non-cardiovascular disease" hinders the observation of the primary outcome "death from cardiovascular disease" and the onset of the two events are correlated. Analysis methods for the competing risk data have been studied extensively in the past two decades. Two main approaches have been investigated: sub-distribution hazard models (Fine and Gray, 1999) and cause-specific hazard models (Gaynor et al., 1993). The sub-distribution proportional hazard model allows direct estimation of marginal effect of covariate on cumulative incidence function (CIF) when different types of events are correlated. In contrast, the cause-specific

proportional hazard model assumes that the different types of events are independent and its marginal probability function describes the event time distribution in a hypothetical situation where no competing events are assumed to occur (Crowder, 2001; Prentice et al., 1978). In this article, we study the variable selection method based on the sub-distribution hazards model in that the sub-distribution hazard model provides the valid estimation that reflects the clinical research reality where a particular covariate may affect various correlated competing risks (Dignam et al., 2012; Fine and Gray, 1999; Lau et al., 2009).

In clinical studies it is common to have non-linear effect of covariate in modeling, which may not be identified if the covariate effect is assumed to be linear. For example, the log-relative hazard function of age may have a convex shape (Kalbfleisch and Prentice, 2002). The linear assumption of covariate effect may be too rigid and unrealistic to depict the covariate effect on hazard. As a result, the proportional hazard models involving splines or local kernel method have been studied by many authors (Fan et al., 1997; Yu and Lin, 2008) in that basis expansion and regression spline methods are popular nonparametric techniques used to characterize nonlinear covariate effects (Gray, 1992; O'Sullivan, 1988, 1993). However, when the underlying covariate effect is linear or zero (no effect), using a regression spline often complicates the interpretation of analysis results. Therefore, investigators often have to trade off between balancing model parsimony and flexibility when using nonparametric covariate functions. It is of interest to develop a data-driven approach to select the covariate functional form for the competing risks model.

In this article, we aim to develop a structure discovery procedure to select linear and non-linear covariate effect in competing risks setting using spectral decomposition (Wand and Ormerod, 2008) and adaptive LASSO (Zou, 2006). The development of variable selection procedures, such as the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) and the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), has been centering around the penalized methods. See Fan and Lv (2010) for a comprehensive review on the popular penalty functions. Recently, variable selection of linear effect has been studied when the competing risks data is independent (Kuk and Varadhan, 2013) or correlated (Ha et al., 2014). We propose the linear/non-linear variable selection by decomposing each covariate into linear and non-linear parts characterized by a set of B-spline bases coefficients which are further treated as two parts. The first part captures the linear effect, whereas the second part captures the non-linear effect. We propose to select non-zero parts of each covariate by applying the adaptive lasso (for the linear part) and group lasso approach (Yuan and Lin, 2006) (for the non-linear part).

The remainder of the article is organized as follows: In Section 2.2 we present the sub-distribution hazard models with potentially nonparametric additive covariate function. In Section 2.3 we present the structure discovery method and the estimation procedure, in which we use a likelihood with the adaptive LASSO penalty term and penalized spline technique in sub-distribution hazard model. Simulation studies and practical examples are presented in Section 2.4 and Section 2.5 respectively. Finally, a brief discussion is given in Section 2.6.

2.2 Sub-distribution hazards model with competing risks data

Let T_i and C_i be the failure and censoring times for the subject i $(i = 1, \ldots, n)$. Let $\Delta_i = I\{T_i \leq C_i\}$ be the indicator of failure or censoring, $\epsilon_i \in \{1, 2, \ldots, J\}$ be the cause of failure of the ith subject if $\Delta_i = 1$, and $\boldsymbol{X_i} \in \mathcal{R}^{1\times p}$ be the covariate vector. For each subject, we denote the observed survival time $T_i^* = min\{T_i, C_i\}$. Assume that $\{T_i^*, \Delta_i, \Delta_i\epsilon_i, \boldsymbol{X}_i\}$ are independent and identically distributed for $i = 1, ..., n$. The cumulative incidence function (CIF), or sub-distribution, for the jth event is defined as $F_j(t) = Pr(T \leq t, \epsilon = j)$. At each time point, the J CIFs are additive to the probability of failure from any cause, or $1 - S(t)$. Throughout the article we denote $j = 1$ as the primary cause of failure. The corresponding sub-distribution hazard defined by Fine and Gray Fine and Gray (1999) is:

$$\begin{aligned}
\lambda_1(t) &= \lim_{\Delta t\to 0} \frac{1}{\Delta t} Pr\left\{t \leq T \leq t + \Delta t, \epsilon = 1 | T \geq t \cup (T < t \cap \epsilon \neq 1)\right\} \\
&= \frac{\partial F_1(t)/\partial t}{1 - F_1(t)} \\
&= -\partial log(1 - F_1(t))/\partial t.
\end{aligned}$$

Note the risk set associated to $\lambda_1(t)$ differs from the traditional risk set in Cox model, as it includes subjects who have not experienced any failure events, as well as those who experienced one of the alternative events (i.e. $\epsilon_i \geq 2$) by the time t. Fine and Gray (1999) developed a proportional sub-distribution hazards model which facilitate the direct estimation of the covariate effects on CIF without estimating the individual cause-specific hazard for different event types.

In order to accommodate both linear and non-linear effects that are associated with the sub-distribution hazard of the primary event, we assume a non-parametric model for the primary competing risk setting:

$$\lambda_1(t; \boldsymbol{X}_i) = \lambda_{10}(t) \exp\left\{\sum_{h=1}^{p} f_h(\boldsymbol{X}_{i,h})\right\}, \tag{2.1}$$

where $\boldsymbol{X}_{i,h}$ is the hth element in the covariate vector, $f_h(\cdot)$ is an unknown covariate function. Baseline sub-distribution hazard $\lambda_{10}(t)$s are assumed to be arbitrary. For the competing risks data, Belot et al. (2010) proposed that the baseline sub-distribution hazard function can be modelled non-parametrically using smoothing cubic splines. In addition, Feng et al. (2005) and Ding and Wang (2008) have shown that a piece-wise constant baseline hazard can perform equally well. However, to the best of our knowledge no work has been done for the selection and estimation of nonparametric covariate function in competing risk models.

Let $f(\cdot)$ denote the unknown function of the covariate. The log-likelihood for the first event derived based on the Fine-Gray model is:

$$l = \sum_{i=1}^{n} I(\Delta_i \epsilon_i = 1) \left\{ \sum_{h=1}^{p} f_h(\boldsymbol{X}_{i,h}) - log\left\{ \sum_{j \in R_{t(i)}} w_j(t_i) \exp\left[\sum_{h=1}^{p} f_h(\boldsymbol{X}_{j,h})\right]\right\}\right\}, \tag{2.2}$$

where $R_{t(i)} = \{j : (T_j \geq T_i) \cup (T_j \leq T_i \cap \epsilon_j > 1)\}$ is the risk set at the time t_i for the ith subject, and $w_j(t_i) = \widehat{G}(t_i)/\widehat{G}(min(t_i, T_j))$ is the weight of the jth at-risk subject from $R_{t(i)}$. Here, $\widehat{G}$ is the Kaplan-Meier estimator Kaplan and Meier (1958) of the survival function of the censoring times ($G(t) = P(C \geq t)$). As described by Fine

and Gray (1999), subjects with $\epsilon_j \geq 2$ should remain in the risk set that contributes to the primary event at time t_i as long as $t_j < C_j$. However, the status of such a subject at time T_j is obviously unknown after time t_i, and is estimated by $w_j(t_i)$. Hence, (2.2) can be viewed as a weighted partial log-likelihood where the weight is defined by the contribution of risk set between subjects who was censored and subjects who experienced competing risks with a weight, and the weight is expressed as the inverse probability of censoring weighting technique introduced by Robins (1993).

To estimate the nonparametric covariate functions, we use penalized spline approach with cubic spline basis. For given set of K inner spline knots, the cubic spline basis functions $\boldsymbol{B}_1, ..., \boldsymbol{B}_{K+4}$ for $f_h(\boldsymbol{X}_h)$ can be generated accordingly using fda package in R Ramsay et al. (2015). Denote the $n \times (K+4)$ design matrix as $\boldsymbol{B}_h = \{\boldsymbol{B}_1(\boldsymbol{X}_h), ..., B_{K+4}(\boldsymbol{X}_h)\}$, with $B(\cdot)$ being the basis function, for $h = 1, \cdots, p$. Also let $\theta_h = (\theta_{1,h}, \theta_{2,h}, ..., \theta_{K+4,h})^T$ be the corresponding B-spline regression coefficients. Plugging the θ_h into (2.2), we can write the penalized likelihood function for estimating $f_h(\cdot)$ as:

$$l_o(\theta_1, \cdots, \theta_p) = \sum_{i=1}^{n} I(\Delta_i \epsilon_i = 1) \left\{ \sum_{h=1}^{p} \boldsymbol{B}_{ih}\theta_h - log \left\{ \sum_{j \in R_{t(i)}} w_j(t_i) \exp \left[\sum_{h=1}^{p} \boldsymbol{B}_{jh}\theta_h \right] \right\} \right\}, \tag{2.3}$$

where $\boldsymbol{B}_{ih}$ is the ith row of $\boldsymbol{B}_h$. To control for the sparsity of parameter estimates, a penalized likelihood with a quadratic penalty can be constructed as:

$$pl_o(\theta_1, \cdots, \theta_p) = l_o(\theta_1, \cdots, \theta_p) - \lambda \sum_{h=1}^{p} \theta_h^T \Omega_h \theta_h, \tag{2.4}$$

where $\boldsymbol{\Omega}$ is the penalty matrix with each kk' entry $\boldsymbol{\Omega}_{kk'} = \int (B_k^{(2)}(s) B_{k'}^{(2)}(s)) ds$ and $\boldsymbol{\lambda}$ is the tuning parameter controlling the smoothness. The penalized likelihood employs a large number of parameters some of which are not needed since some of the continuous covariates may affect the risk in a linear fashion. To determine whether the covariate effects are linear or not, it is desirable to decompose the spline into the linear and non-linear components and apply the variable selection on each components respectively.

2.3 Linear and non-linear variables selection

In this section, we present the decomposition and the selection method based on the penalized likelihood modified from (2.4).

2.3.1 Decomposition of Spline

Following Wand and Ormerod (2008), we apply the spectral decomposition to decompose each covariate function using cubic B-spline basis, $\boldsymbol{B_h}$. Specifically, the penalty matrix can be represented as $\boldsymbol{\Omega_h} = \boldsymbol{U}_h diag(d_h) \boldsymbol{U}_h^T$, with $\boldsymbol{U}_h \boldsymbol{U}_h^T = \boldsymbol{I}$. Matrix $\boldsymbol{U}_h$ consists of column eigenvectors and vector d_h consists of eigenvalues arranged in descending order. Let $d_h = (d_{h+}^T, d_{h0}^T)^T$, where d_{h+}^T is the vector of $K+2$ descending positive eigenvalues, and d_{h0}^T is the vector of two zero eigenvalues. Let $U_h = [U_{h+}, U_{h0}]$ with dimension of $(K+4) \times (K+2)$ and $(K+4) \times 2$ respectively. As Ding and Wang (2008) has shown, we can reparametrize the spline function $f_h(\cdot) = B_h U_h U_h^T \theta$ as

follows:

$$\begin{aligned}
\boldsymbol{B}_h\theta &= \boldsymbol{B}_h\boldsymbol{U}_h\boldsymbol{U}_h^T\theta \\
&= \boldsymbol{B}[\boldsymbol{U}_{h0}\boldsymbol{U}_{h0}^T\theta + \boldsymbol{U}_{h+}diag(d_{h+}^{-1/2})diag(d_{h+}^{1/2})\boldsymbol{U}_{h+}^T\theta_h] \\
&= \boldsymbol{B}_h[\boldsymbol{U}_{h0}\beta_h + \boldsymbol{U}_{h+}diag(d_{h+}^{-1/2})\gamma_h] \\
&= \boldsymbol{C}_h\beta_h + \boldsymbol{Z}_h\gamma_h,
\end{aligned}$$

where $\boldsymbol{C}_h = \boldsymbol{B}_h\boldsymbol{U}_0, \beta_h = \boldsymbol{U}_{h0}^T\theta_h, \boldsymbol{Z}_h = \boldsymbol{B}_h diag(d_{h+}^{-1/2})$, and $\gamma_h = diag(d_{h+})^{1/2}\boldsymbol{U}_{h+}^T\theta_h$. The penalty term in (2.4) can be written as:

$$\begin{aligned}
\boldsymbol{\theta_h^T\Omega_h\theta_h} &= \boldsymbol{\theta_h^T U diag(d_h)U_h^T\theta_h} \\
&= \boldsymbol{\theta_h^T U_{h0} diag(d_{h0})U_h^T\theta_h + \theta_h^T U_{h+} diag(d_{h+})U_{h+}^T\theta_h} \\
&= \gamma_h^T\gamma_h,
\end{aligned}$$

With the reparameterization, the subdistribution hazard model (2.1) can be re-written as:

$$\lambda_1(t; C, Z) = \lambda_{10}(t)\exp\left\{\sum_{h=1}^{p}(\boldsymbol{C}_h\beta_h + \boldsymbol{Z}_h\gamma_h)\right\}. \tag{2.5}$$

Here we plug (2.5) into the penalized likelihood (2.4) with β and γ as the parameters. As suggested by Ding and Wang (2008), in (2.5), $\boldsymbol{C}_h$ is the basis of linear effect and $\boldsymbol{Z}_h$ is the basis of non-linear effect. Since $\boldsymbol{C}$ represents the linear space, we can set $\boldsymbol{C} = [1, \boldsymbol{X}]$ as suggested by Speed (1991). Since the intercept of nonparametric function is not identifiable, we remove the constant vector $\mathbf{1}$ from the $\boldsymbol{C}_h$. The penalized likelihood (2.4) can be re-written as smoothing spline analysis of variance

(SSANOVA) type penalized likelihood

$$pl_{SSANOVA} = pl_o(\zeta) - \lambda_0 \sum_{h=1}^{p} \gamma_h^T \gamma_h, \tag{2.6}$$

where γ_h is the coefficients of the $\boldsymbol{Z}_h$ in (2.5), and λ_0 is a tuning parameter to control the smoothness of the fitted curves. To perform selection of functional covariate form, instead of using these quadratic penalty term, we use the linear and/or nonlinear components for each covariate function by applying the penalty terms allowing for sparse estimation.

2.3.2 Model Selection using Penalized Likelihood

To construct a penalized likelihood function to perform the variable selection, we employ the penalty term adaptive to the feature of each estimated coefficient since zero components are expected to be penalized more and nonzero components are expected to be penalized less. In this way, large nonzero components are protectively preserved in the selection process, while small components are shrunk toward zero. One approach is the adaptive LASSO proposed by Zou (2006), which enjoys the oracle properties by utilizing the adaptively weighted L_1 penalty. To select the linear and non-linear component respectively, we replace the penalty term in (2.6) by the adaptive LASSO penalty terms for linear and non-linear components separately as:

$$pl_{SD}(\zeta) = \frac{1}{n} l_o(\zeta) - \lambda_\beta \sum_{h=1}^{p} \kappa_\beta(\beta_h) - \lambda_\gamma \sum_{h=1}^{p} \kappa_{\gamma_h}(\gamma_h) \tag{2.7}$$

which is a function of $\zeta = (\beta_h, \gamma_h)^T$. For the linear coefficient, the penalty term is $\kappa_\beta(\beta_h) = w_{\beta h}|\beta_h|$, with w_{β_h} denoted as corresponding positive weight (Zou, 2006) for penalty β_h. The weight $w_{\beta_h} = |\tilde{\beta}_h|^{-q}$, where $\tilde{\beta}_h$ is the initial estimator by maximization of (2.6) and q is a positive integer to adjust for the shrinkage of $\kappa_\beta(\beta_h)$. For non-linear components, following Yuan and Lin (2006), we use the grouped LASSO type penalty $\kappa_{\gamma_h} = w_{\gamma h}\|\gamma_h\|$, where $\|\gamma_h\| = (\gamma_h^T\gamma_h)^{1/2}$ is the L_2 norm of γ_h for the hth non-linear component. We choose the weight for the non-linear effect as $w_{\gamma_h} = \|\tilde{\gamma}_h\|^{-r}$, where similar to $|\tilde{\beta}_h|$, $\tilde{\gamma}_h$ is the initial estimator from the maximization of (2.6) and r is a positive number to adjust for the shrinkage of $\kappa_\beta(\beta_h)$. We assume $\tilde{\beta}$ and $\tilde{\gamma}$ are consistent estimators of β and γ respectively (Wand and Ormerod, 2008). Two tuning parameters λ_1 and λ_2 are used to control overall shrinkage imposed on linear and non-linear terms. In this article, we use $q = r = 4$ following the recommendation in Zhang et al. (2011).

By maximizing (2.7), we achieve the linear/non-linear variable selection by shrinking the zero effect to exact zero. Specifically, for linear effect coefficient, an estimate of exact zero is obtained if the true effect is close to zero since $|\beta_h|$ is singular at zero value. For non-linear effect, since a zero non-linear effect is equivalent to each entry of γ_h close to zero and further equivalent to the L_2 norm of γ_h close to zero, the group LASSO penalty assures the sparsity of non-linear coefficients estimated in terms of group estimator of coefficients, according to Zou (2006) study on the oracle property of adaptive LASSO. If both linear and non-linear components are estimated to be zero, we determine the covariate effect as zero effect. If both linear and non-linear components are estimated to be non-zero, we determine the covari-

ate effect as partial-linear effect. If one of the linear or non-linear components is estimated as non-zero while another component is zero, we determine the covariate effect as the corresponding non-zero part effect. Note that there does not exist a continuous second order partial derivatives of (2.7) with respect to β due to the L_1 norm feature of the penalty term. Some special care is required before we apply the Newton-Raphson algorithm. Following Fan and Li (2002), we use the local quadratic approximation (LQA) technique to overcome the difficulty in solving the score equation which has the non-differentiable term at origin and does not have the continuous second-order derivatives. Specifically, we approximate $\psi(|\beta|)$, the L_1 norm penalty term, by quadratic functions as follows. Given an initial value β_0 bounded away from zero, $\psi(|\beta|)$ can be locally approximated by $\{\psi(\beta_0)^2/\beta_0\}\beta^2$, as long as the initial value β_0 is close to the minimizer. In our practice, we choose the estimator from (2.6) as the initial value of estimator and its performance is generally good. To be specific, the derivative of penalty function $\kappa_\beta(\beta)$ can be locally approximated by a quadratic function given an initial value of $\beta_k^{(0)}$ close to the true value of β_k:

$$[\kappa_\beta(|\beta_k|)]' = \kappa'_\beta(\beta_k)sgn(|\beta_k|) \approx \frac{\kappa'_\beta\left(\left|\beta_k^{(0)}\right|\right)}{\left|\beta_k^{(0)}\right|}\beta_k \quad for\ \beta_k \approx \beta_k^{(0)}.$$

By maximizing the penalized log-likelihood in (2.7), we perform variable selection and parameter estimation simultaneously. Specifically, those zero variables (zero linear and non-linear components for each variable) are estimated as zero when the L_1 norm penalty term is used, and they are removed from the model automatically. To

this end, we apply the Newton-Raphson method to estimate (β, γ) by treating each component as the fixed effect. The estimation of the parameters can be obtained by solving the joint estimating score equations of β and γ (without loss of generality, let β_k denote the kth entry of the parameter vector β and $x^{(k)}$ as the corresponding kth component among the pool of all the decomposed components). The score equation of the linear part is:

$$\begin{aligned}\frac{\partial pl_{SD}(\beta,\gamma|\epsilon_i=1)}{\partial \beta_k} &= \frac{\partial l_o(\zeta|\epsilon_i=1)}{\partial \beta_k} - \frac{\partial \kappa_\beta(\beta)}{\partial \beta_k} \\ &= \sum_{i=1}^{n} I(\epsilon_i=1)\left[x_i^{(k)} - \frac{x_i^{(k)} w_j(t_i)\exp\left(\boldsymbol{C}\beta+\boldsymbol{Z}\gamma\right)}{\sum_{j\in R_{t(i)}} w_j(t_i)\exp\left(\boldsymbol{C}\beta+\boldsymbol{Z}\gamma\right)}\right] \\ &\quad - \lambda_\beta|\tilde{\beta}|^{-q}(|\beta_k^{(0)}|)^{-1}\beta_k,\end{aligned}$$

and the score equation of the non-linear part is :

$$\begin{aligned}\frac{\partial pl_{SD}(\beta,\gamma|\epsilon_i=1)}{\partial \gamma_k} &= \frac{\partial l_o(\zeta|\epsilon_i=1)}{\partial \gamma_k} - \frac{\partial \kappa_\gamma(\gamma)}{\partial \gamma_k} \\ &= \sum_{i=1}^{n} I(\epsilon_i=1)\left[x_i^{(k)} - \frac{x_i^{(k)} w_j(t_i)\exp\left(\boldsymbol{C}\beta+\boldsymbol{Z}\gamma\right)}{\sum_{j\in R_{t(i)}} w_j(t_i)\exp\left(\boldsymbol{C}\beta+\boldsymbol{Z}\gamma\right)}\right] \\ &\quad - \lambda_\gamma(\tilde{\gamma}^T\tilde{\gamma})^{-r/2}({\gamma^{(0)}}^T\gamma^{(0)})^{-1}\gamma_k.\end{aligned}$$

To solve the above equations, we can use the standard R package such as "rootSolve".

The standard errors for estimated coefficients can be directly obtained as we are estimating parameters and selecting variables at the same time. Without loss of generality, we denote the parameters (estimated coefficients) as θ in this section. Fine and Gray (1999) proposed to obtain a robust sandwich covariance estimator in that

the martingale properties no longer hold due to the use of IPCW. On the other hand, according to Tibshirani (1996), the variance-covariance matrix can be approximated by the ridge regression of the penalty term, i.e. $\sum_j |\theta_j|$ replaced by $\sum_j \theta_j^2/|\theta_j|$. Therefore we propose the sandwich type variance-covariance matrix estimator. Define the gradient vector $\nabla l(\theta) = \partial l(\theta)/\partial\theta$ and the Hessian matrix $\nabla^2 l(\theta) = \partial^2 l(\theta)/\partial\theta\partial\theta^T$. Following Zhang and Lu (2007), we define $\boldsymbol{A} = diag\{|\hat{\theta}_1|^{-1}, \ldots, |\hat{\theta}_d|^{-1}\}$ and $\boldsymbol{D} = diag\{I\{\theta_1 \neq 0\}|\hat{\theta}_1|^{-1}, \ldots, I\{\theta_d \neq 0\}|\hat{\theta}_d|^{-1}\}$. The sandwich estimate can be calculated as :

$$\widehat{cov}(\hat{\theta}) = (\nabla^2 l(\hat{\theta}) + \lambda \boldsymbol{A})^{-1} \widehat{cov}(\nabla pl(\hat{\theta}))(\nabla^2 l(\hat{\theta}) + \lambda \boldsymbol{A})^{-1}, \tag{2.8}$$

where $\widehat{cov}(\nabla pl(\hat{\theta}))$ can be calculated empirically Fan and Li (2002), or can be estimated by $(\nabla^2 l(\hat{\theta}) + \lambda \boldsymbol{D})[\nabla^2 l(\hat{\theta})]^{-1}(\nabla^2 l(\hat{\theta}) + \lambda \boldsymbol{D})$ Zhang and Lu (2007). In this article we calculate $\widehat{cov}(\nabla pl(\hat{\theta}))$ using the latter approach. Note if an estimated parameter is zero, the corresponding estimated standard error of that parameter is zero as well.

2.3.3 Tuning Parameter Selection and Two-stage Estimation

Variable selection using penalized likelihood approaches highly depends on an appropriate choice of the tuning parameters. The tuning parameters λ_1 for linear components and λ_2 for non-linear components can be selected simultaneously. In variable selection setting, the generalized cross-validation (GCV) statistic has been extensively used Androulakis et al. (2012); Fan and Li (2001, 2002). However, Wang et al. (2007) showed that GCV approach tended to choose the tuning parameters that lead to an overfitted model. Following Wang et al. (2007), Ha et al. (2014) and He et al. (2014), we use a BIC (Schwarz, 1978) type criterion based on the penalized log-likelihood for

tuning parameter selection:

$$BIC(\boldsymbol{\lambda}) = -2l_o(\hat{\boldsymbol{\zeta}}) + log(n) \times df_{\boldsymbol{\lambda}}, \tag{2.9}$$

where $\hat{\zeta}$ are the estimators obtained by maximizing (2.7) at a given tuning parameter combination, denoted as $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$, and $l_o(\hat{\boldsymbol{\zeta}})$ is the value of (2.2) evaluated at the estimated $\hat{\boldsymbol{\zeta}}$. $\boldsymbol{\lambda}$ is determined by a brutal search of a grid of possible values and is selected as the one that minimizes $BIC_{\boldsymbol{\lambda}}$. The $df_{\boldsymbol{\lambda}}$ is the total number of non-zero estimates of $\hat{\boldsymbol{\zeta}}$.

In order to reduce the estimation bias due to the L_1 penalty term, we suggest to use a two-stage process proposed by Zhang et al. (2011). In the first stage, the penalized likelihood is maximised to select potentially non-zero linear and/or nonlinear covariate functions. Conditional on for selected model from first stage, we perform a regular parameter estimation based on the SSANOVA type of penalized likelihood in the second stage. Since the structure has been discovered in the first stage, only variables selected in the first stage are used for maximization in the second stage. We consider our procedure is valid based on the work of Zhang et al. (2011), which proved the consistency of the two stage structure discovery procedure and proved the convergence rate of the first stage estimation from SSANOVA in modeling using normal distributed data. Berk et al. (2013) systematically discussed the post-selection inference validity for linear effects modeling, and suggested that it is feasible to achieve the valid inference that does not depend on selection of the true model by framing the analysis in the context of simultaneous inference. It is shown that, for coefficient

estimation of linear effect model, the post-selection inference error can be controlled by simultaneous inference, i.e. taking into account the multiplicity associated with all linear functions of estimates. Although it is not a trivial topic on extending from linear effect to non-linear effect estimation scenario, such as non-parametric spline estimation, we do not pursue this investigation further in this topic.

2.4 Simulation Study

We conduct simulation studies, based on 300 replications of simulated data set (Fan and Li, 2002; He et al., 2014), to evaluate the performance of our proposed method. We generate n observations data set under various scenarios. For each subject within each data set, two competing events are considered with $\epsilon_i = 1$ representing the primary event. Various approaches could be used to generate the competing risks data (Beyersmann et al., 2009), and we use the approach employed in Fine-Gray. Specifically, let $\mathbf{X}$ and η_1 denote the covariate vector and corresponding covariate function for the primary event. Five covariates $(x_1, x_2, x_3, x_4, x_5)$ are generated from the uniform distribution $Uniform(0, \pi)$ and the correlation matrix of five covariates was assumed to be compound symmetric with all the diagonal entries 1 and off-diagonal entries r, where r varies as 0,0.3 and 0.7 as 3 scenarios of random datasets generation. A parameter p related to the probability of experiencing the primary event. Also p is used to adjust the ratio between two types of events. In this article, we set $p = 1$ and obtain a 3:1 ratio of the two events. For the ith subject, the cause of failure of primary event ϵ_i was simulated as: $Pr(\epsilon_i = 1|\eta_1) = 1 - (1 - p)^{exp(\eta_1)}$. Conditional on the cause, the subdistribution of

type 1 event given x_i was $F_1(t|\eta_1) = P(T \leq t, \epsilon = 1|\eta_1) = 1 - [1 - p(1 - e^{-t})]^{exp(\eta_1)}$, where $p = P(\epsilon_i = 1|\mathbf{X} = \mathbf{0})$ specified a mixture of a unit exponential and a degenerated random variable with mass $1 - p$ at ∞, and $\eta_1 = \sum_{h=1}^{p} f_h(x_h)$, with $f_h(x_h)$ be the true effect of the hth covariate.

The conditional distribution function of T_i given a primary event as well as $\mathbf{X}$ is:

$$F(t|\mathbf{X}, \epsilon = 1) = \frac{1 - [1 - p(1 - e^{-t})]^{exp(\eta_1)}}{1 - (1 - p)^{exp(\eta_1)}}. \tag{2.10}$$

For the ith subject, the failure time conditional on cause 1 is obtained by generating y_i as a uniformly distributed variable on interval $(0, 1-(1-p)^{exp(\eta_i)})$, and then applying the inverse function of $F_1^{-1}(t|\eta_{1_i}, \epsilon_i = 1) = -log(1 - p^{-1}(1 - (1 - y_i))^{exp(-\eta_{1_i})}))$. Conditional on cause 2, the failure time for the ith subject was generated from the exponential distribution with the rate $\exp\{\eta_{2_i}\}$, where η_{2_i} was chosen as $-\eta_{1_i}$. So the distribution function for the competing event is:

$$F(t|\mathbf{X}, \epsilon = 2) = 1 - \exp\left\{-e^{\eta_2} t\right\}. \tag{2.11}$$

Finally, Censoring times are generated from a $Uniform(0, c)$ distribution where the value of c is chosen to obtain the desired censoring rate. We use the censoring rate 10% and two sample sizes, $n = 250$ and 500.

The true subdistribution hazard expressed by the five covariates is:

$$\lambda_{1_i}^{sub}(t) = \lambda_{01_i}^{sub}(t) \exp\left\{0x_{1_i} + x_{2_i} + sin(x_{3_i}) + (x_{4_i} - 1)^2 + 0x_{5_i}\right\} \tag{2.12}$$

where $\lambda_{01_i}^{sub}(t) = \frac{pe^{-t}}{1-p[1-e^{-t}]}$. The covariate functions of x_1, x_2, x_3, x_4, x_5 are zero, linear, non-linear, partial linear and zero, respectively. Three scenarios of different pairwise correlation coefficients are 0, 0.3, and 0.7, respectively. In each scenario, 100 random data sets were generated and the proposed method is applied to discover the structure feature of covariates in the first stage and refit the model using the selected covariate function forms in the second stage. The tuning parameters $\lambda_1,\ \lambda_2$ are determined by minimizing the BIC criterion as defined in (2.9).

The simulation results are summarized in the Tables 2.1 and 2.2 and Figure 2.1. Table 2.1 reports the percentage of correct selection of both linear and non-linear features of each covariate effect over 300 random data sets, named as "selection accuracy" here, for various sample sizes and correlation scenarios. When sample size is 250 and independent ($r = 0$) scenario, among the 5 covariates, the selection accuracy for the linear effect is 87%, for the pure non-linear effect is 94%, and for the partial linear effect is 100%. For the two zero-effect covariates the prediction accuracy are both more than 95%. The selection accuracy for both pure linear and pure non-linear effect covariates decreases as the correlation among the covariates increased as expected, especially when the correlation coefficient is high (greater than 0.5). On the other hand, the selection accuracy for zero and partial linear effect remains high even for the settings with higher correlation. The same sign correlation leads to more than 10 times overestimation of true coefficient (Yoo et al., 2014), which affects mainly for non-linear coefficients. Since the prediction accuracy criterion is based on correctly selecting both linear and non-linear parts, the overestimated coefficients tends to misspecify the pure linear and pure non-linear effects as partial linear which lead

Table 2.1: Summary of Variables Selection Accuracy (%)

		Noise	Non-Zero Effect			Noise
		0	X_2	$sin(X_3)$	$(X_4-1)^2$	0
		Z [a]	L [a]	NL [a]	PL [a]	Z [a]
$n=250$						
	$r=0$	97	87	94	100	96
	$r=0.3$	96	86	92	100	92
	$r=0.7$	96	80	86	100	90
$n=500$						
	$r=0$	99	92	97	100	98
	$r=0.3$	99	86	95	100	98
	$r=0.7$	98	83	90	100	96

[a] Z: zero effect; L: linear effect; NL: non-linear effect; PL: partial linear effect.

to a higher false positive rate due to the overestimation in non-linear components. We also conduct a simulation with a larger sample size of 500 subjects with the same parameter setting, which is presented in the Table 2.1 also. The selection accuracy of all the five covariates are improved compared with smaller scenario.

After the selection from first stage, we perform second stage estimation. Table 2.2 summarises averaged integrated squared error (AISE) of the fitted curves for all the selected non-zero effect variables after the two-stage estimation procedure. We see the AISE increases as the correlation among variables increases, but declines as the sample size increases.

Figure 2.1 shows the estimation performance for all the 3 scenarios. Figure 1(a-c) show estimators of non-linear and partial linear effects respectively when $r = 0$. The fitted curves (gray) are generally close to the true curve (black solid curve). The

Table 2.2: Summary of AISE of the Selected Non-zero Variables

		X_2	$sin(X_3)$	$(X_4-1)^2$
		L [a]	NL [a]	PL [a]
$n=250$				
	$r=0$	0.0416	0.2479	0.2313
	$r=0.3$	0.0878	0.2684	0.2358
	$r=0.7$	1.1385	0.3421	0.3564
$n=500$				
	$r=0$	0.0216	0.2302	0.2276
	$r=0.3$	0.0378	0.2413	0.2258
	$r=0.7$	1.0685	0.3421	0.3104

[a] L: linear effect; NL: non-linear effect; PL: partial linear effect.

corresponding 95% point-wise confidence interval based on estimated standard error (black dash) are very close to the 95% point-wise confidence interval based on empirical standard deviation (black dotted curve). Figure 1(d-i) show the estimated curves for the $r = 0.3$ and $r = 0.7$. Generally, the biases are small in the point estimates and the standard error. The bias increases slightly as the correlation coefficient increases, but remains small. On the other hand, estimated SE and empirical SD are generally very close over the domain of both effects for various correlation scenarios. Overall, estimated standard error for the three non-zero covariates are close to the 95% nominal level empirical standard deviation and the estimation biases are small.

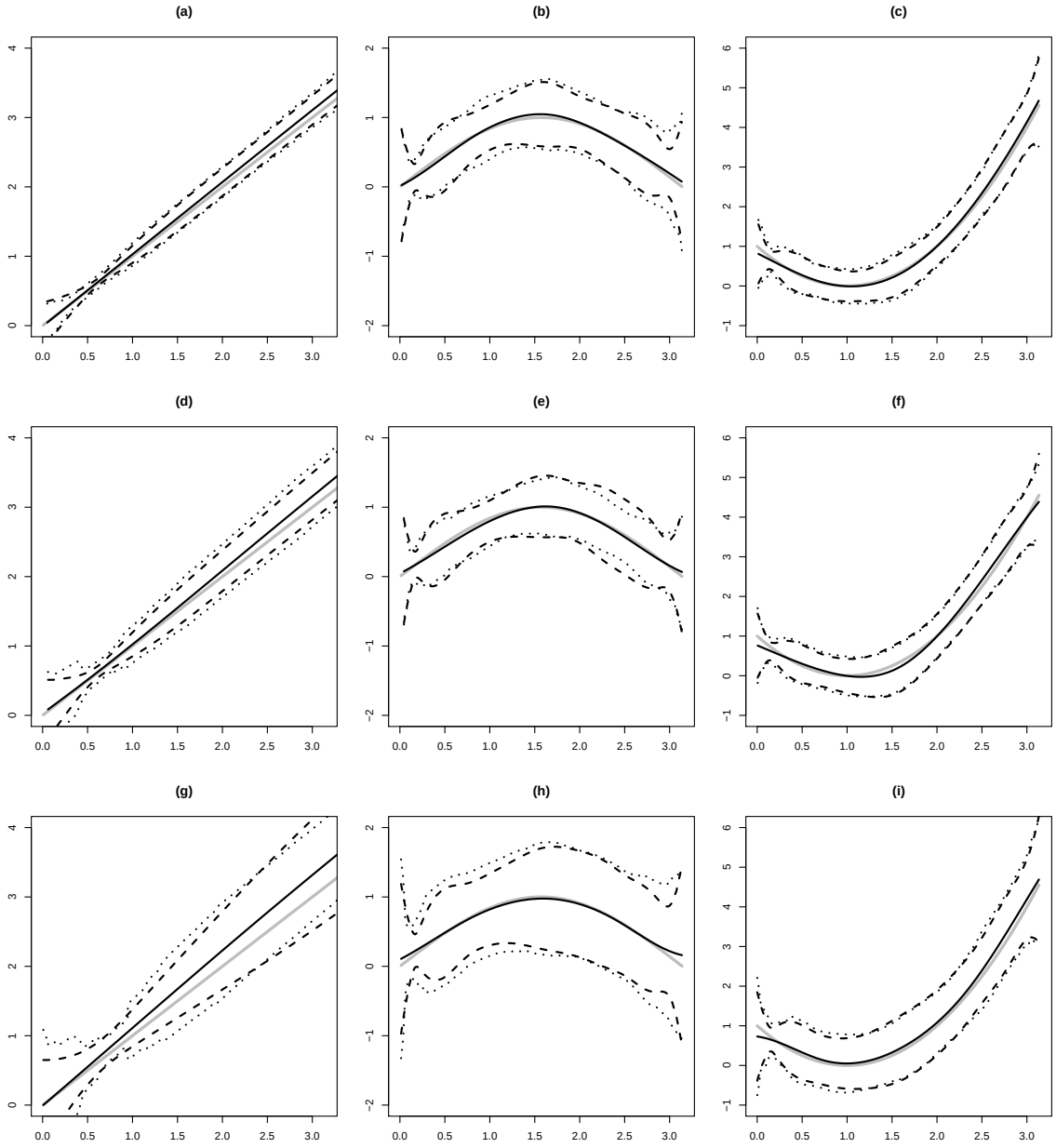

Figure 2.1: Estimated curves of the three non-zero effect covariates from 100 replicates. The plots are, from the top to the bottom, for r=0, 0.3 and 0.7 respectively. In each row, the plots from the left to the right are linear, non-linear and partial linear effects. The gray lines are the true curves. The solid black lines are the average of 300 replicates. The dashed lines are the 95% confidence intervals based on the estimated standard error. The dotted lines are the 95% confidence intervals based on the empirical standard deviation.

2.5 Practical Examples

We apply the proposed structure discovery procedure to analyze the risk factors of cardiovascular and non-cardiovascular death using the Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) Study conducted between 1998 and 2002 (The AFFIRM Investigators, 2002, 2004). AFFIRM was a randomized, multi-center clinical trial comparing treatment strategies designed for atrial fibrillation (AF) in patients who were 65 years old or above at enrolment or who had other risk factors for either stroke or death. Patients were randomized to either the rhythm-control or the rate-control treatment strategy group. A total of 4060 patients with AF were recruited. Among them, 279 (6.87%) had experienced Non-cardiovascular death and 329 (8.10%) Cardiovascular death. The The AFFIRM Investigators (2002) concluded that there was no difference in all-cause mortality between the two treatments. Steinberg et al. (2004) re-analyzed the AFFIRM data focusing on cause-specific mortality and concluded non-cardiovascular mortality was significantly associated with rhythm-control treatment while cardiovascular mortality shows no significant association with rhythm-control treatment.

We are interested in evaluating the potential linear and non-linear risk factors on the cardiovascular mortality using the proposed selection and estimation procedure. We apply the nonparametric selection and estimation method to the AFFIRM data. Following Steinberg et al. (2004), we include a total of 12 categorical covariates and 8 continuous covariates collected at baseline: Minority, Female, LVD (Left Ventricular Dysfunction), CHF (History of Congestive heart failure), Hypertension, Stroke,

Fast (Fast heart rate during AF in last six months), Duration (Duration of qualifying episode of atrial fibrillation), NYHC (Current Congestive heart failure status), First (First episode of atrial fibrillation), CAD (History of Coronary artery disease), Arm (assigned as Rhythm control treatment at time of randomization), Rate (Heart rate), Age (Age at Baseline), MaxVR (Maximum recorded ventricular rate of qualifying episode of AF), CurVR (Current Ventricular heart rate), BPSys (Systolic blood pressure), BPDias (Diastolic blood pressure) and Days (Number of inpatient days for critical care) and CHAscore (A score combining a list of cardiovascular clinical measure). All the interaction terms of baseline categorical variables with treatment are included in the initial model for variable selection, and the continuous variables are selected using our two-stage linear and non-linear variable selection procedure. When assuming the cardiovascular death as the primary event, we have covariates including Arm, LVD, CHF, CAD, Stroke, NYHC, Age, CurVR and Days selected as non-zero risk factors in the first stage. The estimation result of all non-zero main effect and two-way treatment-covariate interaction effects are summarized in Tables 2.3 and 2.4 and Figure 2.2. We apply the two-stage linear/non-linear variable selection and estimation procedure on the AFFIRM data set. In Table 3, we report the baseline characteristic factors that can be used as predictor for both causes of death. The treatment strategy(rhythm control) has significant interaction with gender and hypertension status. Therefore, we evaluate the treatment effect in each gender and hypertensive status groups. In Table 2.4, one can see that for the female and hypertensive patients, the rhythem control group has a marginally significant higher cardiovascular death risk (45.7% higher hazard, P-value 0.081) than those being treated by rate control. For males without hypertension, the rhythm control group has a significant lower risk of

Table 2.3: Summary of effect estimate of discrete variables by modelling cardiovascular death as primary failure

Variable	Hazard Ratio	Coefficient	Standard Error	P-value
Rhythm Control	0.4891	-0.7151	0.3122	0.0219
LVD	1.4643	0.3814	0.0656	$< .0001$
CHF	1.7591	0.5648	0.1621	0.0005
CAD	1.4572	0.3765	0.1592	0.0097
Stroke	1.9893	0.6878	0.1472	$< .0001$
NYHC	1.1965	0.1794	0.0877	0.0382
Rythm*Female	1.7189	0.5417	0.2682	0.0438
Rythm*Hypertension	1.7333	0.5500	0.3291	0.0950

Table 2.4: Summary of subgroup rhythm control effect by modelling cardiovascular death as primary failure

Subgroup	Hazard Ratio	Coefficient	Standard Error	P-value
Female and Hypert.	1.4574	0.3765	0.21557	0.0811
Female and Non-Hypert.	0.8408	-0.1734	0.33781	0.6077
Male and Hypert.	0.8478	-0.1651	0.18334	0.3680
Male and Non-Hypert.	0.4891	-0.7151	0.31220	0.0219

cardiovascular death risk than the rate control group. Figure 2.2 shows the fitted curve of non-linear effect. The risk of cardiovascular death increase sharply as the enrollment age increase from 50 to 55 years and then reach a plateau after age 55 years. The risk also increases as the current ventricular rate increases in general. The critical inpatient days after randomization is in proportional to a higher hazard in general.

We also conduct the linear/non-linear variable selection for the non-cardiovascular modeling and report the results in Section 2.5. There is no significant treatment covariate interaction term selected in the non-cardiovascular death modeling.

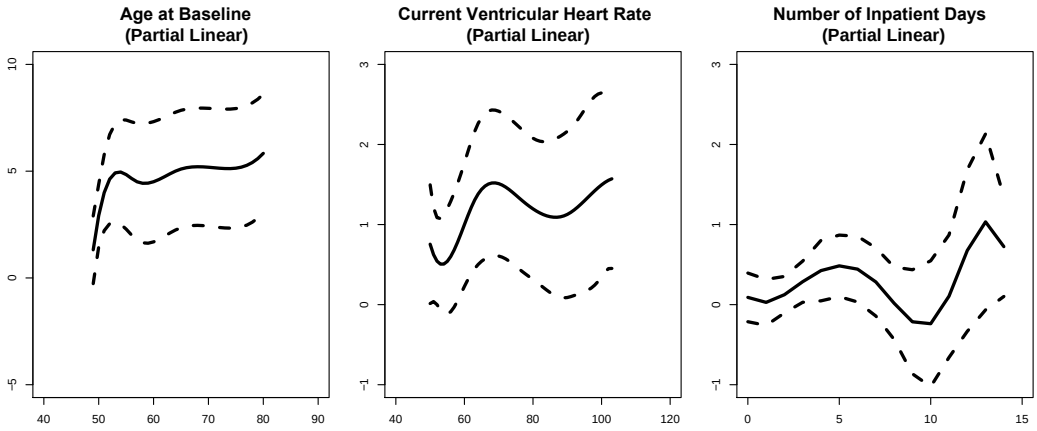

Figure 2.2: Fitted curves of the selected non-zero continuous covariates by modelling cardiovascular with the estimated 95% confidence intervals.

Table 2.5: Summary of effect estimate of discrete variables by modelling non-cardiovascular death as primary failure

Variable	Hazard Ratio	Coefficient	Standard Error	P-value
Rhythm Control	1.5981	0.4688	0.1441	$< .0001$
Female	0.5318	-0.6315	0.1585	0.0001
CHF	1.3998	0.3364	0.1607	0.0142
Durat	1.0688	0.0666	0.0530	0.2108
Age	1.0757	0.0730	0.0199	$< .0001$

2.6 Discussion

We propose a linear/non-linear variable selection method in the Fine-Gray sub-distribution proportional hazards model setting. The model extends the structure discovery method from linear model (Zhang et al., 2011) to the competing risks sub-distribution hazards model setting and the simulation results show good performance of selection and estimation. By applying the variable selection method such as adaptive LASSO onto the linear and non-linear components of the spline function, we determine the covariate functional form (zero/linear/nonlinear). Also by jointly applying penalized likelihood methods and nonparametric smooth splines method, we use our method to identify the true effect of variables and build the connection between variable selection and non-parametric estimation of sub-distribution hazard model. This connection provides more flexibility for variable screening and variable effects estimation. An important advantage of our proposed method is that maximization of our penalized likelihood function is readily applicable in most major statistical softwares. Using this approach for the real data analysis, our method provides an alternative to illustrate the functional feature of covariate effect. Our method is a data-driven approach for determining the functional covariate form in the competing risks data analysis. Additionally, according to the simulation, the number of inner knots has very little impact on the accuracy of model selection. As a result, for complicated models one could use a smaller number, say three, of inner knots to enhance computational efficiency in the first stage, and then increase the number of inner knots in the second stage to achieve desired estimation accuracy. Compared with the reported performance in variable selection in frailty model (Ha et al., 2014),

variable selection in joint model (He et al., 2014), variable selection in Cox model (Fan and Li, 2002), adaptive LASSO in Cox model (Zhang and Lu, 2007) and linear and non-linear variable selection in linear regression (Zhang et al., 2011), we consider the proposed structure discovery procedure performs well.

In this article we assume the number of variables, p, is less than sample size n. Future studies should address the issue of extending the variable selection method to the high dimensional data. Another potential future research direction could be selection of time-varying coefficient in Fine-Gray model. Similar to Belot et al. (2010) work, our study assume the time independent covariate effect. Yan and Huang (2012) proposed the model selection for Cox models with time-varying coefficients. It is promising that the structure discovery tool can be used to overcome this limitation mentioned in their paper.

Chapter 3

Estimation of Time-dependent ROC Curves with Interval Censored Survival Data in Competing Risks Setting

3.1 Introduction

In recent years, extending the traditional time-independent (cross-sectional) concept of the capacity for discriminating diseased from non-diseased subjects to accommodate the additional dimension of time has been an area of active research (Cai et al., 2006; Heagerty et al., 2000; Zheng et al., 2010; Zheng and Heagerty, 2007). See Pepe et al. (2008) for a review on such extensions. In this setting, disease status is generally considered as a function of the disease onset at time t, $D(t)$, and accuracy summaries are time-dependent functions. Following the tradition of medical diagnostic research, there are two main approaches to describe the accuracy of a marker Z measured at baseline: the retrospective and prospective summaries of accuracy. Two commonly used retrospective measures for quantifying the accuracy of Z in predicting $D(t)$ are the time-dependent true positive rate (TPR) and false positive rate (FPR), respectively, defined as

$$TPR(c;t) = P\{Z \leq c \mid D(t) = 1\},$$

$$FPR(c;t) = P\{Z \leq c \mid D(t) = 0\}.$$

The receiver operating characteristics (ROC) curve evaluated at a FPR of v, $ROC(v;t) = TPR[FPR^{-1}(v;t);t]$, is often used to summarize the time-dependent trade-off between TPR and FPR with varying cut-off values in a common scale for comparing the accuracy of the biomarker, which may not be measured in same magnitude, in distinguishing the diseased/non-diseased subjects at given time.

The analysis of time-to-event data is complicated by the presence of interval censoring and competing events, both of which occur frequently in clinical studies. Time-to-event data may be subject to both interval censoring (for time) and competing risks (for event). Interval censored data arise when an exact failure time can not be observed, but can only be determined to lie within an interval. Competing risks data arise when study subjects are at risk of more than one types of events of interest, and the occurrence of one event may prevent the occurrence of other potential events, thus only the earliest event can be observed. Sometimes competing risks data may not have all the events recorded in exact time format. For example, the ADNI study have the death recorded as exact time (date), but the primary event, onset of dementia, is recorded in an interval censored format (three months window). Such an issue need to be addressed by accommodating the interval censoring feature in the competing risks analysis.

When we study the association between failure and covariates, it is of interest to model the cause-specific cumulative incidence function (CIF) (Dignam et al., 2012; Lau et al., 2009). There are typically two ways to semi-parametrically model CIF based on profile likelihood: Proportional hazards model and proportional odds model.

Li (2016) used the estimation approach by Zhang et al. (2010) to the problem of semi-parametric regression of the CIF with interval-censored competing risks data under the proportional subdistribution hazards assumption of the Fine-Gray model (Fine and Gray, 1999). However, the assumption of proportional subdistribution hazards may not be correct and the interpretation of the subdistribution hazards is difficult (Fine, 2001). Therefore Fine and Gray (1999) and Eriksson et al. (2015) proposed to overcome the drawback by modeling CIF with covariates to facilitate the estimation of marginal effect from covariates. Fine and Gray (1999) modeled sub-distribution hazard of primary event under a proportional hazard model and built the connection between CIF and covariate effect. However, the validity of the model relies on the assumption of proportional sub-distribution hazards, and Fine (2001) pointed out that the parameters in proportional sub-distribution hazards model are quite difficult to interpret. To relax such an assumption, Eriksson et al. (2015) proposed a proportional odds cumulative incidence model for competing risks data for the sake of simple and useful interpretation of the regression parameters. Recently, Bakoyannis et al. (2016) extended Zhang et al. (2010) and Li (2016)'s method into a class of semi-parametric generalized odds ratio transformation models using the B-spline sieve maximum likelihood estimation of the cause-specific cumulative incidence function for interval censored data.

Statistical procedures for estimating ROC functions under the competing risks setting are well developed. Recently, Saha and Heagerty (2010) defined causes-specific ROC curves by stratifying cases by event types and proposed estimation procedures for these quantities. Their method extends the time-dependent ROC function es-

timation to competing risks setting using nonparametric (Kaplan-Meier estimator and nearest neighbor estimator) and semi-parametric (maximum partial likelihood estimation) approaches, based on the cause-specific hazards. To extend Saha and Heagerty (2010) approach to incorporate covariates, Zheng et al. (2012) proposed a semi-parametric approach to estimate the cause-specified ROC functions based on two definitions of cases and controls. Their method provides a practical approach to estimate the time-dependent ROC functions in competing risks setting based on the right censored time-to-event data. It is reported the cause-specific hazards have limitation in clinical research application, as the censoring view of any other competing events may hamper the interpretation of the covariate's effect on primary event (e.g. complication for some primary event in certain disease) (Dignam et al., 2012; Fine, 2001; Lau et al., 2009). In this article, we adopt Bakoyannis et al. (2016) proportional odds model for interval censored competing risks CIF estimation method and Song and Zhou (2008) ROC function estimation to extend the right-censored competing risks time-dependent ROC function estimation to the interval-censored scenario. The purpose of this article is to provide a theoretical justification and numerical study validation of applicability of the method. In theory, we prove the consistency of cumulative/incident ROC functions estimation in the interval censored competing risks survival setting. In application, we explore the convergence rate by numerical methods to show the plausibility of estimation efficiency and confirm the consistency of our proposed method.

The rest of the article is organized as follows. In Section 3.2, we review the parameter estimation and its corresponding asymptotic large sample property of the

proportional odds model (Bakoyannis et al., 2016), and derive the explicit form the resulting ROC functions. We also provide the proof of consistency of the proposed estimators. In Section 3.3, we conduct simulation studies to assess the performance of our method, and compare the estimation consistency and robustness of the proposed approach. In Section 3.4, we apply our method to estimate the ROC functions of the biomarker from the Alzheimer's disease neuroimaging initiative (ADNI) study as an illustration. We conclude with a brief discussion in Section 3.5, and outline the technical proofs in Appendices.

3.2 Method

3.2.1 Background of the Interval Censored Data

Suppose that

$$0 < Y_{i,1} < Y_{i,2} < \cdots < Y_{i,n_i} < \infty$$

are ordered examination times for the ith subject,$i = 1, \ldots, n$. Denote $Y_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,n_i})$. Let T_i be the ith patient's true failure time. Computationally, the interval censoring can be considered in three possibilities: (i) failure occurs before the first examination time. Denote $U_i = Y_{i,1}$ and let $V_i = Y_{i,2}$. Let $\delta_{1,i} = 1_{[T_i \leq U_i]}$ and $\delta_{2,i} = 1_{[U_i < T_i \leq V_i]}$. Then $\delta_{1,i} = 1$ and $\delta_{2,i} = 0$. (ii) T_i is known to be bracketed between a pair of examination times $(Y_{i,L}, Y_{i,R})$, where $Y_{i,L}$ is the last examination time preceding T_i and $Y_{i,R}$ is the first examination time following T_i. Denote $U_i = Y_{i,L}$ and $V_i = Y_{i,R}$. as in (i). Then $\delta_{1,i} = 0$ and $\delta_{2,i} = 1$. (iii) At the last examination, the

failure did not occur. Then $\delta_{1,i} = 0$ and $\delta_{2,i} = 0$. The effective observations are

$$(\delta_{1,i}, \delta_{2,i}, U_i, V_i), \qquad i = 1, \ldots, n.$$

Estimation of the Cox proportional hazards model and the proportional odds regression model was considered by Huang and Wellner (1997) and Huang and Rossini (1997), respectively. They showed that the MLEs of the regression parameters in both models are asymptotically normal and efficient, even though the MLEs of the baseline cumulative hazard function or odds function only have $n^{1/3}$-rates of convergence.

Following Bakoyannis et al. (2016), throughout this article we assume the following basic assumptions:

(A1) The (unobservable) failure time is independent of the examination times given the covariates.

(A2) The joint distribution of the examination times and the covariates are independent of the parameters of interest.

The cause-specific CIF is defined as

$$F_j(t) = P(T \leq t, \delta = j), \qquad j = 1, 2, \ldots, J.$$

Here, for each j these functions are increasing functions of t. In order to overcome the shortcoming that the separately estimated J CIFs may be sum up to a value greater than 1 (Choi and Huang, 2014), Bakoyannis et al. (2016) imposes an additional

constraint, over all the covariate patterns (denoted as z), to ensure that the sum of the estimated CIFs at the maximum follow-up time is bounded above by 1 as follows:

$$\max_{z} \left\{ \sum_{j=1}^{J} F_j(t; z) < 1 \right\}.$$

For the interval censored data with a single event, the joint density can be expressed using vector as $X = (\delta^{(1)}, \delta^{(2)}, \delta^{(3)}, U, V, Z)$, where $\delta^{(s)} \in \{0, 1\}$ for $s = 1, 2, 3$ and $\delta^{(1)} + \delta^{(2)} + \delta^{(3)} = 1$, is

$$p(x) = F(u|z)^{\delta^{(1)}}[F(v|z) - F(u|z)]^{\delta^{(2)}}(1 - F(v|z))^{\delta^{(3)}} h(u, v, z),$$

where $h(\cdot)$ is the joint density function of (U, V, Z). The log-likelihood function of an independent sample $(\delta_i^{(1)}, \delta_i^{(2)}, \delta_i^{(3)}, U_i, V_i, Z_i), i = 1, \ldots, n$ with the same distribution as $(\delta^{(1)}, \delta^{(2)}, \delta^{(3)}, U, V, Z)$ is

$$l_n = \sum_{i=1}^{n} \{\delta_i^{(1)} log F(U_i|Z_i) + \delta_i^{(2)} log[F(V_i|Z_i) - F(U_i|Z_i)] + \delta_i^{(3)} log(1 - F(V_i|Z_i))\}. \quad (3.1)$$

When we extend the single event interval censored survival data to settings with J competing risks, we can extend the notations as follows. Let $j = 1, ..., J$ denotes a number of competing risks. If a subject fails from the jth cause of failure before the first examination time U, we observe $\delta_j^{(1)} = 1$. If this subject fails between U and V, we observe $\delta_j^{(2)} = 1$. If this subject is right censored (i.e. $T > V$), we observe $\sum_{j=1}^{J}(\delta_j^{(1)} + \delta_j^{(2)}) = 0$. We assume the observations interval be $[a, b]$ be $a \leq U < V \leq b$.

Along with $T, C, U, V, \delta_j^{(1)}$ and $\delta_j^{(2)}$, we also observe a vector of biomarker $Z \in \mathbb{R}^d$ with coefficient β which is the primary parameter of interest. In this article, we assume $d = 1$, i.e. a single variable or a composite score of several variables.

3.2.2 Proportional odds model for CIF with interval censored data

Recently, Eriksson et al. (2015) proposed the approach to directly model CIF for right censored data using a proportional odds model. Therefore, the CIF can be straightforwardly used to estimate the ROC functions under Song and Zhou (2008) Cumulative/Incident ROC functions framework. Proportional odds models have been rigorously studied for univariate survival data (Bennett, 1983; Chen et al., 2002; Murphy et al., 1997). With the CIF of cause $j(j = 1, ..., J)$ denoted as $F_j(t|\ Z)$, it is the special case of the general class of semi-parametric transformation models $g(F_j(t|\ Z)) = H_j(t) + Z^T\beta_j$, where $H_j(t)$ is an unspecified positive monotone increasing function and $g(\cdot)$ is a known increasing link function. In this article, we consider the proportional odds model for the CIF

$$logit(F_j(t|Z)) = logit(F_j(t; \beta, H, Z)) = log(H_j(t)) + Z^T\beta_j,$$

where $H(t)$ is an increasing positive function with $H(0) = 0$. The cumulative incidence is thus linear on the logit scale with intercept, logarithm of baseline hazard, increasing over time and time-constant log-odds ratio for failure from cause j. The

model implies that the CIF of cause j is

$$F_j(t|Z) = \frac{H_j(t)exp(Z^T\beta_j)}{1 + H(t)exp(Z^T\beta_j)}, \tag{3.2}$$

and the subdistribution hazard of failure from cause j can be derived from

$$-\frac{\partial}{\partial t}log(1 - F_j(t;\ Z)) = \left[exp\{-Z^T\beta_j\} + H_j(t)\right]^{-1} \frac{\partial H_j(t)}{\partial t}.$$

Bakoyannis et al. (2016) extended the proportional odds model to the interval censored data scenario using sieve maximum likelihood estimation based on B-spline. Therefore this model can avoid specific parameter assumption of baseline cumulative incidence function $\Phi(t)$. The density of one observation in with

$$F_j(t|Z) = \frac{exp(\Phi_j(t) + \beta_j Z)}{1 + exp(\Phi_j(t) + \beta_j Z)}, \tag{3.3}$$

where β is the regression parameter for biomarker Z, and $\phi(t) = log(F(t)/(1-F(t)))$ is the baseline monotone increasing log-odds function. The maximum likelihood estimator is the $(\hat{\phi}, \hat{\beta})$ that maximizes $l_n(\alpha, \beta)$ under the constraint that $\hat{\phi}$ is a nondecreasing function and $\sum_{j=1}^{J} F_j(t) < 1$.

3.2.3 Definition and asymptotic properties of ROC function

For survival data there are several potential extensions of cross-sectional sensitivity and specificity. Rather than a simple binary outcome, $Y_i = 1$ a survival time can be viewed as a time-varying binary outcome by focusing on the counting process representation $N_i(t) = I\{T_i \leq t\}$. Accuracy extensions are classified according to whether the "cases" used to define time-dependent sensitivity are *incident* cases where $T = t$, or equivalently $dN_i(t) = 1$, is used to define cases for time t, or *cumulative* cases where $T < 1$ or $N_i(t) = 1$ is used. We also consider whether "controls" are *static*, defined as subjects with $T_i > t^*$ for a fixed value of t^*, or whether controls are *dynamic* and defined for time t as those subjects with $T_i > t$. Following Heagerty and Zheng (2005) we use the superscripts $\mathbb{C}$ and $\mathbb{I}$ to denote cumulative and incident definitions of sensitivity, and use the superscripts $\mathbb{D}$ and to denote definition of dynamic specificity.

For a baseline marker value, Heagerty et al. (2000) proposed versions of time-dependent sensitivity and specificity using the definitions as

$$sensitivity^{\mathbb{C}}(z;t) : P(Z_i > z|T_i \leq t) = P(Z_i > z|N_i(t) = 1)$$

$$specificity^{\mathbb{D}}(z;t) : P(Z_i \leq z|T_i > t) = P(Z_i \leq z|N_i(t) = 0).$$

Based on these definitions, the entire population is classified as either a diseased case ($I\{T_i \leq t\}$) or a non-diseased control ($I\{T_i > t\}$) by time t on the basis of the dichotomized diagnostic test result ($Z_i > z$) at time t. Also, the ith individual plays the role of a control for times $t < T_i$, but then contributes as a case for later times,

$t \geq T_i$. ROC curves are defined as $ROC^{\mathbb{C}/\mathbb{D}}(p;t) = TPR^{\mathbb{C}}\left\{[FPR^{\mathbb{D}}]^{-1}(p;t)\right\}$ where $TPR^{\mathbb{C}}(c;t) = P(X_i > c \mid N_i(t) = 1)$ and $FP_t^{\mathbb{D}}(c;t) = P(X_i > c \mid N_i(t) = 0)$. In the absence of censoring, $ROC^{\mathbb{C}/\mathbb{D}}(p;t)$ can be estimated using the empirical distribution of the marker separately among cases and controls. With censored survival times Heagerty et al. (2000) developed a non-parametric estimator based on either Kaplan-Meier estimator or the nearest-neighbor bivariate application estimator of Akritas (1994). On the other hand, Heagerty and Zheng (2005) proposed the *incident/dynamic* definitions of sensitivity and specificity adopting Heagerty et al. (2000) and Etzioni et al. (1999) as

$$sensitivity^{\mathbb{I}}(c;t) : P(X_i > c \mid T_i \leq t) = P(X_i > c \mid dN_i^*(t) = 1)$$
$$specificity^{\mathbb{D}}(c;t) : P(X_i \leq c \mid T_i > t) = P(X_i \leq c \mid N_i^*(t) = 0).$$

where $dN_i^*(t) = N_i(t) - N_i(t-)$. Using this definition, each subject does not change disease status and is treated as either a case or a control. Cases are stratified according to the time at which the event occurs (incident) and controls are defined as those subjects who are event free through a fixed follow-up period, $(0, t)$ (static).

3.2.4 Estimation of ROC function with interval censored competing risks data

In this section we establish the notation. Let T_i denote the actual failure time for subject $i, (i = 1, ..., n)$. We assume each subject may experience J mutually exclusive types of causes of failure and denote the cause for the ith subject as $\epsilon_i = j, (j = 0, 1, ..., J)$, where $\epsilon_i = 0$ denotes the right censoring. Let (U_i, V_i) de-

note the observed event time interval where the ith subject experienced the observed event. And we also observe a biomarker covariate Z (here for simplicity we assume Z is one dimension). The log-likelihood function (3.1) of the data in terms of CIF, $F_j(t)$, can be built as the expression of $(U_i, V_i, \epsilon_i, Z)$. Here, the proportional odds model of $F_j(t)$ is suggested to model the covariate effects, as (3.3) expressed, where $\Phi_j(t)$ is a nonparametric function to express the baseline odds of CIF as (3.3). So for any subject the CIF can be expressed by parameters $\Phi(t)$ and β, and β is time independent.

The ROC function of competing risks data involves the stratification of cases. Following Saha and Heagerty (2010) we consider two causes of failure for simplicity: $\epsilon_i = 1, 2$. Here we consider a single control group as controls are free of any event and cases may be accrue due to either one of the two event types. The stratified cases with common controls are defined as:

$$\begin{aligned} &Case\ 1\ : T \leq t,\ \epsilon = 1; \\ &Case\ 2\ : T \leq t,\ \epsilon = 2; \\ &Control : T > t \qquad . \end{aligned}$$

We can estimate the cumulative ROC function based on the cumulative TPR for event types 1 and 2, and the FPR. A cumulative ROC curve for each event type can

be obtained by plotting the cause-specific cumulative TPR versus the common FPR:

$$TPR_1^C(c,t) = P(Z > c \mid T \leq t, \epsilon = 1),$$
$$TPR_2^C(c,t) = P(Z > c \mid T \leq t, \epsilon = 2),$$
$$FPR(c,t) \ = P(Z > c \mid T > t, \epsilon = 0).$$

Similarly, an incident ROC curve for each event type can be obtained by plotting the cause-specific cumulative TPR versus the common FPR:

$$TPR_1^I(c,t) = P(Z > c \mid T = t, \epsilon = 1),$$
$$TPR_2^I(c,t) = P(Z > c \mid T = t, \epsilon = 2),$$
$$FPR(c,t) \ = P(Z > c \mid T > t, \epsilon = 0).$$

These ROC curves measure the predictive accuracy of the marker to distinguish among subjects who experience a particular type of event by/at time t and those who do not experience any event by time t. A marker that is selected to seek the subjects who are likely to experience a particular event is expected to have a high sensitivity to detect these cases, while it may be less sensitive at identifying those subjects who die from other events.

Song and Zhou (2008) proposed the estimation of cumulative and incident ROC curves adjusting covariate effect and showed the advantage over Heagerty-Lumley-Pepe and Heagerty-Zheng's approaches in terms of efficiency of estimators using simulation

studies. The Song-Zhou ROC function estimations are

$$\widehat{FPR}(z;t) = \frac{\int_{z}^{\infty} \hat{S}(t|u)d\hat{P}(Z \le u)}{\int_{-\infty}^{\infty} \hat{S}(t|u)d\hat{P}(Z \le u)} = \frac{\sum_{i=1}^{n} \hat{S}\{t|Z_i\}I(Z_i \ge z)}{\sum_{i=1}^{n} \hat{S}\{t|Z_i\}},$$

$$\widehat{TPR}_C(z;t) = \frac{\int_{z}^{\infty} \{1-\hat{S}(t|u)\}d\hat{P}(Z \le u)}{\int_{-\infty}^{\infty} \{1-\hat{S}(t|u)\}d\hat{P}(Z \le u)} = \frac{\sum_{i=1}^{n} \left[1-\hat{S}\{t|Z_i\}\right] I(Z_i \ge z)}{\sum_{i=1}^{n} \left[1-\hat{S}\{t|Z_i\}\right]},$$

$$\widehat{TPR}_I(z;t) = \frac{\int_{z}^{\infty} \hat{f}(t|u)d\hat{P}(Z \le u)}{\int_{-\infty}^{\infty} \hat{f}(t|u)d\hat{P}(Z \le u)} = \frac{\sum_{i=1}^{n} \hat{f}(t|Z_i)I(Z_i \ge z)}{\sum_{i=1}^{n} \hat{f}(t|Z_i)},$$

where $\hat{f}(t) = \partial\hat{S}(t)/\partial t$.

We develop the corresponding quantities following Song and Zhou's concepts. With some algebra, for the first event the cumulative true positive function (denote as $TPR_1^C(z;t)$), incident true positive function (denote as $TPR_1^I(z;t)$) and false positive function (denote as $FPR(z;t)$) can be estimated as:

$$\widehat{FPR}(z;t) = \frac{\sum_{i=1}^{n}((1-\hat{F}_1(t|Z_i)-\hat{F}_2(t|Z_i)) * I\{Z_i > z\}}{\sum_{i=1}^{n}(1-\hat{F}_1(t|Z_i)-\hat{F}_2(t|Z_i))}, \tag{3.4}$$

$$\widehat{TPR}_C^{(1)}(z;t) = \frac{\sum_{i=1}^{n} \hat{F}_1(t|Z_i) * I\{Z_i > z\}}{\sum_{i=1}^{n} \hat{F}_1(t|Z_i)}, \tag{3.5}$$

$$\widehat{TPR}_I^{(1)}(z;t) = \frac{\sum_{i=1}^{n} \hat{F}_1(t|Z_i) * (1+exp\{\hat{\phi}(t)+\hat{\beta}Z_i\})^{-1} * I\{Z_i > z\}}{\sum_{i=1}^{n} \hat{F}_1(t|Z_i) * (1+exp\{\hat{\phi}(t)+\hat{\beta}Z_i\})^{-1}}, \tag{3.6}$$

where $\partial \hat{F}_1(t)/\partial t = \hat{F}_1(t|Z_i) * (1 + exp\{\phi(t) + \beta Z_i\})^{-1}$.

Thus the estimators of $ROC^C(v;t)$ and $ROC^I(v;t)$ for the event 1 are $ROC_C^{(1)}(v;t) = TPR_C^{(1)}[FPR^{-1}(v;t);t]$ and $ROC_I^{(1)}(v;t) = TPR_I^{(1)}[FPR^{-1}(v;t);t]$, respectively. For valid estimation, t should be less than the maximum follow-up time. In Appendix we provide the proof of the consistency of the true positive function and false positive function estimators.

In order to compare our method with ROC estimation based on the estimators from the modeling that treats interval-censored data as right-censored data, i. e. only the right terminal of the interval will be used as the presumable event time. In this article we compare our approach with the proportional odds model (Eriksson et al., 2015) approach that treats interval censored data as right censored. We conduct the simulation with cumulative and incident ROC functions under same definition of cases/controls with same TPR/FPR quantities estimated using these two modeling respectively and report the results along with that from our proposed modeling. Note in the proportional odds model the covariate effects are connected with CIF, denoted here as $F^{PO}(t)$), using the logit link, as

$$F_j^{PO}(t) = \frac{H_{j0}(t)exp\{\beta_j Z\}}{1 + H_{j0}(t)exp\{\beta_j Z\}}, \qquad j = 1, 2,$$

where $H_{j0}(t)$ is the baseline odds for the jth event with $H_{j0}(0) = 0$. We derive the corresponding cumulative and incident TPR/FPR functions as follows.

$$\widehat{FPR}(z;t) = \frac{\sum_{i=1}^{n}((1 - \hat{F}_1^{PO}(t|Z_i) - \hat{F}_2^{PO}(t|Z_i)) * I\{Z_i > z\}}{\sum_{i=1}^{n}(1 - \hat{F}_1^{PO}(t|Z_i) - \hat{F}_2^{PO}(t|Z_i))}, \tag{3.7}$$

$$\widehat{TPR}_1^C(z;t) = \frac{\sum_{i=1}^{n} \hat{F}_1^{PO}(t|Z_i) * I\{Z_i > z\}}{\sum_{i=1}^{n} \hat{F}_1^{PO}(t|Z_i)}, \tag{3.8}$$

$$\widehat{TPR}_1^I(z;t) = \frac{\sum_{i=1}^{n} exp\{\beta Z_i\} * \hat{F}_1^{PO}(t|Z_i) * I\{Z_i > z\}}{\sum_{i=1}^{n} exp\{\beta Z_i\} * F_1^{PO}(t|Z_i)}, \tag{3.9}$$

where $\partial \hat{F}_1^{PO}(t)/\partial t = exp\{\beta Z_i\} * \hat{F}_1^{PO}(t) * [\partial \hat{H}_{10}(t)/\partial t]$.

3.3 Simulation Study

Extensive simulation studies were conducted to assess the finite sample behavior of the estimators proposed in Section 3.2. We consider various scenarios by varying the sample size of data sets. We conduct a set of simulation studies. We assume two causes of a failure and one biomarker that is correlated with both of the causes since cumulative incidence function is used in the setting where the competing risks (correlated) "share" the covariate effect (Dignam et al., 2012). Following Bakoyannis et al. (2016), we have the CIF for causes 1 and 2 generated from:

$$F_j(t) = \frac{exp\{\phi_j(t) + \beta_j Z\}}{1 + exp\{\phi_j(t) + \beta_j Z\}}, \qquad j = 1, 2,$$

where $exp\{\phi_1(t)\} = 0.4[1-exp(-0.6t)]/0.6$ and $exp\{\phi_2(t)\} = 0.75[1-exp(-0.5t)]/0.5$ follow Gompertz distributions as in Jeong and Fine (2007), and the biomarker Z is generated from the standard normal distribution. The two causes of events are generated from Binomial distribution with probability $exp\{0.67 * \beta_j * Z\}/(1 + exp\{0.67 * \beta_j * Z\})$, where the true parameter β_j is set to be 1 for $j = 1, 2$. The actual event time for the two events are generated respectively from $F_j^{-1}(u)$, where u is a size n random variable from $Uniform(0, 1)$. For each subject the corresponding left and right observed times (U, V) are generated by mapping the two adjacent numbers from a series of sequential exponential distributed random numbers with mean parameter 3. Finally, censoring times are generated from a $Uniform(0, c)$ distribution where the value of c is chosen to obtain a 10% censoring rate. This setting yields a $2 : 3$ ratio for the two events at infinite time, and approximate $3 : 5$ ratio calculated empirically from the generated data. The simulation study results of time-dependent ROC curves from interval censoring survival data in competing risks setting are summarized in Table 3.1 and Table 3.2.

We consider four scenarios as sample size $n = 300, 600, 900$ and 1200. For each scenario, we generate 1000 simulated data sets. For each simulated data set, we estimate the $\mathbb{C}/\mathbb{D}$ and $\mathbb{I}/\mathbb{D}$ ROC functions at time $t = 1$ (denoted as "IC"). For the standard error of the ROC functions, we do the bootstrap method based on 100 resampled data sets to investigate the possible bias of standard error estimation. For each data set generated, we obtain the point estimators of cumulative and incident ROC functions evaluated at 0.1, 0.3, 0.5, 0.7 and 0.9 FPR values and report the difference with true estimator, as well as the corresponding standard error using our proposed method.

We also calculated the sample standard deviation (SD) and the averaged standard error (SE) over 1000 simulations, and the 95% empirical confidence interval coverage percentage (CP) for each estimated ROC function. Additionally, since we do not prove the asymptotical normality of the estimator, we calculate the exponent of the empirical convergence rate of SD for $n = 600, 900$ and 1200 using $n = 300$ as reference from

$$\frac{log(SD_1/SD_2)}{log(n_2/n_1)}.$$

For all the four sample sizes, we can see our method yielded negligible bias. Also We can see the empirical diverges from 0.5 but tends to approach 0.5 as sample size increases. The bootstrap SE estimations from the proposed approaches are close to empirical SD. As the sample size increases we can see the SD and SE get closer which implies the convergence of the estimator though we still lack the proof of large sample property of the estimator. Therefore the empirical coverage probability, which is based on the SE estimated using bootstrap, is close to nominal level 0.95. For comparison purpose, we also calculated the ROC function based on the CIF estimated using proportional odds model (Eriksson et al., 2015) by treating the interval-censored data as right censored (denoted as "RC"), i.e. only the right terminal of each subject's censoring window is used. We report the difference of estimators evaluated at 0.1, 0.3, 0.5, 0.7 and 0.9 FPR values in Table 3.1. We see the estimation on all the cut-off points have negligible bias, and the estimated standard errors are all close to the corresponding empirical standard deviations, and the empirical coverage percentages are close to the nominal level 95% generally. For the naive approach , we can see the

difference is considerable compared with that from the proposed approach, which also leads to lower coverage percentage which is considerably away from the 95% nominal level. The estimation bias reduced as sample size increase. The bias of standard error estimator reduced even more substantially. Overall, our proposed approach works well in terms of small bias and efficiency.

Table 3.1: Simulation results for ROC estimators evaluated at $t = 1$ for $n = 300, 600$ scenarios. Bias is the empirical bias ($\times$ 1000); SD is the empirical standard deviation ($\times$ 1000); SE is the averaged bootstrapping-estimated standard errors ($\times$ 1000); ECR is the exponent of empirical convergence rate.

			Cumulative ROC				Incident ROC			
n	Cut	C	Bias	SD	SE	ECR	Bias	SD	SE	ECR
300										
	0.1	IC	3.98	53.37	51.35	-	3.41	52.25	49.88	-
		RC	-51.29	46.10	43.20	-	-95.62	31.87	30.12	-
	0.3	IC	2.01	48.41	45.23	-	1.83	48.40	45.00	-
		RC	-59.26	46.97	44.66	-	-99.65	40.36	38.37	-
	0.5	IC	0.37	33.12	30.65	-	0.34	33.34	30.77	-
		RC	-48.17	35.41	33.76	-	-75.48	33.39	31.78	-
	0.7	IC	-0.30	17.42	16.16	-	-0.29	16.33	19.55	-
		RC	-30.22	20.95	19.98	-	-44.59	21.07	20.01	-
	0.9	IC	-0.18	4.31	4.16	-	-0.17	4.39	4.23	-
		RC	-9.91	6.29	6.20	-	-13.67	6.71	6.57	-
600										
	0.1	IC	0.23	36.45	35.73	0.501	0.10	35.93	34.67	0.540
		RC	-45.07	29.97	31.06	0.621	-89.03	20.04	21.16	0.669
	0.3	IC	-0.57	32.33	31.94	0.582	-0.56	32.76	31.82	0.563
		RC	-52.42	31.15	31.97	0.592	-92.84	26.32	27.17	0.616
	0.5	IC	-0.88	21.93	21.72	0.594	-0.86	22.43	21.89	0.571
		RC	-42.21	23.29	24.03	0.604	-69.50	21.79	22.53	0.615
	0.7	IC	-0.73	11.67	11.45	0.577	-0.72	12.00	11.63	0.444
		RC	-26.29	13.56	14.06	0.627	-40.58	13.62	14.10	0.629
	0.9	IC	-0.23	2.97	2.93	0.537	-0.23	3.08	3.00	0.511
		RC	-8.66	4.15	4.25	0.599	-12.37	4.47	4.54	0.586

Table 3.2: Simulation results for ROC estimators evaluated at $t = 1$ for $n = 900, 1200$ scenarios. Bias is the empirical bias ($\times$ 1000); SD is the empirical standard deviation ($\times$ 1000); SE is the averaged bootstrapping-estimated standard errors ($\times$ 1000); ECR is the exponent of empirical convergence rate.

			Cumulative ROC				Incident ROC			
N	Cut	C	Bias	SD	SE	ECR	Bias	SD	SE	ECR
900										
	0.1	IC	1.73	27.12	28.98	0.616	1.40	26.23	28.07	0.627
		RC	-46.83	25.71	24.99	0.531	-90.86	17.24	17.10	0.559
	0.3	IC	1.71	24.54	25.78	0.618	1.58	24.71	25.68	0.611
		RC	-52.49	26.68	25.58	0.514	-92.70	22.51	21.79	0.531
	0.5	IC	1.01	16.87	17.55	0.614	0.98	17.22	17.67	0.601
		RC	-42.11	20.57	19.17	0.494	-69.20	19.19	18.01	0.504
	0.7	IC	0.37	9.11	9.18	0.590	0.36	9.38	9.31	0.504
		RC	-26.27	11.83	11.21	0.520	-40.43	11.80	11.26	0.527
	0.9	IC	0.03	2.28	2.32	0.576	0.03	2.38	2.37	0.557
		RC	-8.56	3.50	3.38	0.531	-12.22	3.72	3.62	0.536
1200										
	0.1	IC	3.75	26.09	25.73	0.516	3.10	24.16	24.90	0.556
		RC	-48.38	23.43	22.13	0.488	-92.81	16.32	15.03	0.482
	0.3	IC	3.18	23.25	22.89	0.529	2.89	22.49	22.82	0.552
		RC	-54.80	24.25	22.63	0.476	-95.60	20.79	19.20	0.478
	0.5	IC	1.99	15.44	15.50	0.550	1.89	15.18	15.62	0.567
		RC	-43.52	18.14	16.93	0.482	-71.03	17.11	15.86	0.482
	0.7	IC	0.90	8.06	8.07	0.555	0.88	8.04	8.20	0.511
		RC	-27.24	10.54	9.84	0.495	-41.67	10.63	9.86	0.498
	0.9	IC	0.19	1.99	2.02	0.557	0.19	2.01	2.07	0.563
		RC	-8.83	3.17	2.95	0.494	-12.56	3.41	3.15	0.488

3.4 Data Examples

We provide an illustration with the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The study has been conducted since 2004 and is currently on-going, aiming to improve clinical trials for the prevention and treatment of Alzheimer's disease (AD) (Jones-Davis and Buckholtz, 2015; Mueller et al., 2017; Weiner et al., 2010). We consider two competing events: onset of dementia and death. All the patients were followed up to 96 months after enrollment. In total, 502 dementia-free patients at baseline were included in the study. Among whom 69 (13.7%) patients ending with dementia, which were considered as the primary event, and 41 (8.2%) ended with death, which will be considered as the competing event. ADAS-13 (ADAS-Cog scale based on 13 items at baseline) has been proposed as a prognostic marker according to the literature (Skinner et al., 2012) and the result from the initial variable selection procedure also confirms it. We considered ADAS-13 as the biomarker. It is a crucial first step to evaluate the predictive performance of ADAS-13 based criteria for early screening. In order to uncover the clinical utility of ADAS-13 in guiding intervention decision, it is crucial to evaluate the predictive performance of ADAS-13 based screening. Subjects who are mostly likely to develop dementia could be recommended for aggressive treatment.

We performed model estimation for the dementia with death as the competing event. Proportional odds assumption was checked to be appropriate using method proposed by Eriksson et al. (2015). To compare the accuracy of the ADAS-13 score as the biomarker in distinguishing the subjects developing dementia by a given time t and

those developing dementia progression after t, we estimated the cumulative ROC curve for the ADAS-13 score using the estimator $\widehat{ROC}_C$. In Figure 1, we plot the estimated cumulative ROC curves (solid black curve) for the composite marker at $t = 2,\ 4$ and 8 years post baseline. The 95% point-wise confidence intervals (dashed black curves) are estimated via the bootstrap method using 200 resampled data sets. Since it is common that investigators may neglect the interval censoring nature of the collected data and treat them as right censored data, here we also present the cumulative ROC curves (gray curves) estimated using proportional odds model (Eriksson et al., 2015) as naive approach in the plots for reference.We compare the estimated ROCs based on the estimators from the two approaches at three time points in Figure 1. The black curve corresponds to the proposed proportional odds estimator and the gray curve to the naive approach, with the dashed curves for 95% confidence interval estimated using bootstrap based on 100 re-sampled data sets.

From Figure 1 we can see the naive approach estimated ROC curves lead to larger difference from the proposed approach as time increases. Table 3.3 reports the area under curve (AUC) of the two approaches for 3 time points. The AUC of the estimated cumulative ROC curve for the biomarker ADAS-13 with the corresponding 95% confidence interval using numerical method are presented in Table 3.3. The AUC value over 3 time points all have its lower bounder of confidence interval above 0.5, which indicates the effectiveness of the ADAS-13 as a biomarker to distinct dementia cases and controls at all time points post baseline given death's impact as a competing risk. Generally, the composite biomarker ADAS-13 have AUC value all greater than 0.5 in classifying Alzheimer diseased subjects and non-diseased subjects

Table 3.3: Estimation of area under curve (AUC) for two approaches in biomarker ADAS-13 at $t = 2, 4$ and 8 years after enrollment.

	Method	AUC	95% Confidence Interval
t=2			
	IC	0.649	(0.563, 0.734)
	RC	0.688	(0.654, 0.721)
t=4			
	IC	0.638	(0.560, 0.717)
	RC	0.678	(0.634, 0.711)
t=8			
	IC	0.597	(0.547, 0.648)
	RC	0.665	(0.604, 0.697)

at various time points. The objective of the present analysis is twofold. First we use the proposed marker to evaluate its performance to discriminate between subjects who experienced AD by t years versus those who were dementia free by t years using cumulative ROC function. Secondly, this objective can be interpreted as to identify those subjects who are at "high risk" and for whom intervention is warranted. From the results we see the baseline ADAS-13 score can be used as a tool for early screening of the patients with potential future onset of dementia.

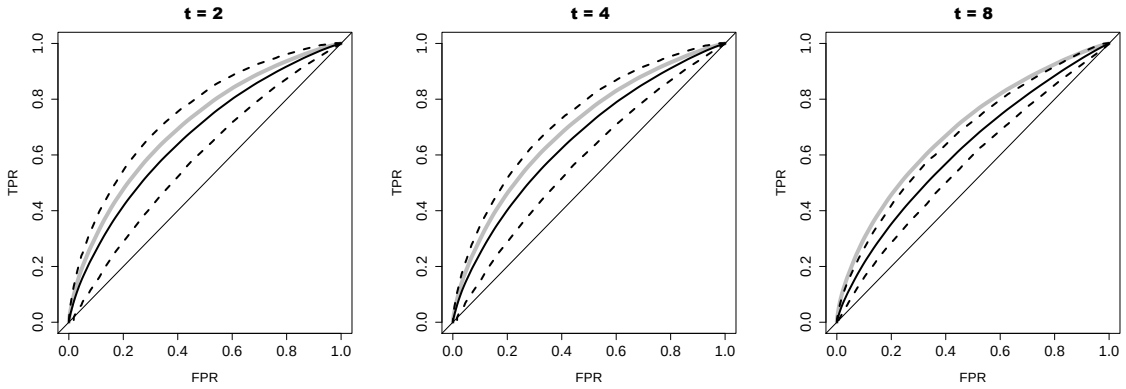

Figure 3.1: Estimated cumulative ROC curves for the ADAS-13 using ADNI data. The plots are, from the left, for t =2, 4 and 8 years post baseline. Cumulative ROC curves estimated using proposed method is indicated by solid black lines. Estimated 95% point-wise confidence intervals corresponding to the proposed approaches are presented as dashed lines. ROC curves estimated from naive approach are indicated by solid gray curve.

3.5 Discussion

In this article, we present ROC function estimation for characterizing time-dependent classification accuracy summaries of a prognostic marker when there are potentially more than one cause of failure and data is subject to interval censoring. To incorporate competing risks, we stratify the cases by causes of failure and controls are a common group of patients who remain free of any events at the prediction time. To incorporate interval censoring, we implement the proportional odds model to jointly estimate the cause-specific CIFs of causes with constraint that the sum over all the CIFs is bounded by one. Though due to the unavailability of the asymptotic normality property of the estimated CIF we can only prove the consistency of the estimator, our simulations indicate that the estimated standard error converges to the empirical standard deviation of the estimators as the sample size increases, which implies the convergency existence but the rate is not $n^{1/2}$. The consistency of estimation is

proved and verified in the simulations, and the consistency depends on the correct specification of the interval censoring nature of the data and it can be greatly impacted by misspecified censoring status as non-negligible bias as is shown in the simulations. Further work is warranted in respect of the property of the AUC.

Chapter 4

Estimation of Time-dependent ROC Curves with Clustered Survival Data

4.1 Introduction

The receiver operating characteristic (ROC) curve is a plot of the true positive rate (TPR) (i.e. probability of identifying a case when the subject is truly diseased) versus false positive rate (FPR) (i.e. probability of identifying a case when the subject is not diseased) at different possible thresholds. accuracy. If Y denotes the diagnostic test or marker, with higher values more indicative of disease, and D is a binary indicator of disease status, then the ROC curve for Y is a plot of the sensitivity associated with the dichotomized test $X > c$ versus (1 - specificity) for all possible threshold values c. Therefore, the ROC function is a monotone function with independent variable $P(Y > c \mid D = 0)$ versus dependent variable $P(Y > c \mid D = 1)$ defined on the domain ranging over critical value $c \in (-\infty, \infty)$, where $sensitivity(c, t) = P\{Y < c \mid D(t) = 1\}$ and $specificity(c, t) = P\{Y \leq c \mid D(t) = 0\}$. An ROC curve provides a graphical characterization of the separation between the two binary outcomes' distributions (diseased versus nondiseased). If the binary outcomes are distinguished completely then the ROC curve takes the value 1 (perfect TPR) for any FPR greater than zero. In this situation the marker is perfect at discriminating between cases and controls. Interpretations of the ROC curve is that the higher the ROC curve is in the quadrant $[0, 1] \times [0, 1]$, the better its capacity is for discriminating diseased from nondiseased

subjects. General discussions of ROC analysis can be referred to Tosteson and Begg (1988), Zweig and Campbell (1993), Pepe and Cai (1993), Pepe et al. (1999) and Pepe (1997, 1998, 2003).

In the past two decades, many researchers have extended the binary outcome ROC to survival data. For example, Heagerty et al. (2000) proposed a nonparametric approach for the time-dependent ROC curve based on the incident TPR and the dynamic FPR, using the Kaplan-Meier estimator of the survival distribution and the empirical distribution estimator of the biomarker. The proposed estimation provided a common scale to compare the prediction accuracy among different markers. Heagerty and Zheng (2005) took a semi-parametric approach for the time-dependent ROC curve based on the cumulative TPR and the dynamic FPR using a proportional hazards model for biomarker variables. Both the Heagerty-Lumley-Pepe and Heagerty-Zheng approaches can be used to evaluate and compare biomarkers in classifying subjects based on their survival times (Heagerty and Zheng, 2005). The former is useful in distinguishing subjects failing by a given time and those failing after this time, and the latter is useful in distinguishing subjects failing at a given time and those failing after this time. Cai et al. (2006) estimated the time-dependent ROC curve based on the cumulative TPR and static FPR, assuming standard binary regression models for the cumulative TPR and the static FPR, and a proportional hazards model for the censored distribution. Song and Zhou (2008) justified the superiority of their proposal to extended Heagerty-Zheng's semi-parametric estimation approach to covariate-specific ROC curves and show the superiority of his proposed estimation in terms of efficiency.

All the above publications studied ROC function estimation for independent survival data. In practice, however, survival data are often collected in clusters, such as paired data from subjects' eyes (non-informative cluster size) in ophthalmology studies, or survival data of various infected organs from cattle (informative cluster size) in a veterinary study, where the existing methods of ROC estimation are no longer valid. Clustered or correlated survival or failure-time data arise when each study subject may experience multiple events or when there exists some natural or artificial clustering of subjects inducing dependence within cluster. Biomedical examples include the sequence of tumor recurrences or infection episodes, the development of physical symptoms or diseases in several organ systems, the occurrence of blindness in the left and right eyes, the onset of a disease among family members. Clustered survival data using both parametric and semi-parametric models have been studied in the past 30 years. To accommodate the correlated structure of the failure times, Wei et al. (1989) proposed to use a working independent partial likelihood of the marginal proportional hazards model for an estimation. The marginal PH model does not impose any assumption on the interdependence among the multivariate failure times and therefore is quite flexible. Extensive studies on marginal proportional hazards model are reported in the literature, and readers may refer to see Wei et al. (1989), Cai and RL (1995),Cai and RL (1997), Gray and Li (2002), Lee et al. (1992), Prentice and Hsu (1997), Pepe and Cai (1993), Hughes (1995), Yang and Ying (2001) and Chen et al. (2010). In particular, the estimation method considered in Wei et al. (1989) is based on a pseudo-likelihood that is a product of marginal partial likelihoods. The method is conceptually clear, numerically simple, and easy to implement. However, the pseudo-likelihood approach does not best capture the interdependence of multivariate failure

times and may not produce the most accurate estimation of regression parameters. In fact, for the MPH model considered in this paper, the pseudo-likelihood estimation can be significantly improved in some cases. There are alternative estimation methods existing, such as weighted partial likelihood score equation (Cai and RL, 1995, 1997; Gray and Li, 2002).

Consequently, most attention over the past two decades has been confined to marginal hazard models and frailty models. The frailty model considers the conditional hazard given the unobservable frailty variables, which is particularly useful when the association of failure types within a subject is of interest (Hougaard, 2000). However, such models tend to be restrictive with respect to the types of dependence that can be modeled and model fitting is usually cumbersome. When the correlation among the observations is unknown or not of interest, the marginal hazard model approach which models the "population-averaged" covariate effects has been widely used. In this article we focus on marginal modeling approach. Liang and Zeger (1986) proposed a class of generalized estimating equations (GEE) methods to handle the dependent repeated data type, and they used the GEE methods on longitudinal data analysis. Wei et al. (1989) applied GEE method in multivariate survival data analysis. However, their method may not work well for if the clustered binary data has different correlation structure and informative cluster size, which occurs when the cluster size is affected by the outcome. To overcome the limitation that the GEE methods may not work well if the clustered data has different correlation structure and informative cluster size (Liang and Zeger, 1986), which occurs when the cluster size is affected by the outcome, Hoffman et al. (2001) proposed a novel within-cluster

resampling (WCR), where one observation is randomly sampled from each cluster. The observations in the resampled dataset are thus independent and the standard methods can be readily applied. By resampling the observed data with replacement many times, we can obtain estimators through averaging over the estimators from the resampled data. Hoffman et al. (2001) showed that WCR method works well with various within-cluster correlations and account the effect of informative cluster size. WCR method is computational intensive but yields consistent and asymptotically normal estimators. It has two advantages over GEE for the analysis of clustered data. First, WCR handles the correlation structure in a fully ad hoc way, so that the correlation structure is not required to be specified. Secondly, WCR remains valid in the presence of informative cluster size, whereas the GEE method does not, because WCR method can account for informative cluster sizes by assigning equal chance when we resample from different clusters. Hoffman et al. (2001) developed the WCR method and applied the method to angular data, Bayesian inference, p-value, vector parameters, genetics data and random cluster sizes. Follmann et al. (2003) established the asymptotic theories and application of the WCR method, referred as multiple outputation. More recently, Miao (2014) applied the WCR method to ROC function estimation for binary data.

In this article, we extend the cumulative and incident ROC function estimation (Song and Zhou, 2008) to clustered survival data setting using WCR of Hoffman et al. (2001) to perform the estimation of ROC in clustered survival data. Our method can be implemented straightforwardly and achieve unbiased and consistent results. The theory of WCR applied in marginal modeling of the clustered survival data is well developed

in the GEE context (Cong et al., 2007), and the asymptotic theory of time-dependent ROC function is developed (Song and Zhou, 2008). We investigate the WCR method for clustered survival data such that the resampled independent data can be analyzed using the conventional Cox model. And we compare our estimated standard error of parameters with that estimated using naive approach, since the estimation equation is identical by marginal and marginal approaches (Wei et al., 1989).

The rest of the article is organized as follows. In Section 4.2, we review the parameter estimation and its corresponding asymptotic large sample property of ROC function in independent survival data. We also review the WCR method and justify its validity in ROC estimation for clustered survival data. In Section 4.3, we conduct simulation studies to assess the performance of our method, and compare the estimation consistency and robustness of the proposed approach. In Section 4.4, we apply our method to a real data analysis as an illustration. We provide some discussion in Section 4.5.

4.2 Method

4.2.1 Marginal estimation based on partial likelihood estimating equations

Let $i = 1, ..., n$ index the clusters which are assumed to be independent of each other, and $k = 1, ..., K$ denote the individuals within each cluster. Let T_{ik} and C_{ik} be the failure and censoring times for the kth individual from the ith cluster, respectively. Let Z denote the biomarker, where $t \in [0, \tau]$ for some finite constant $\tau > 0$. We assume that T_{ik} is conditionally independent of C_{ik} given Z_{ik}. The observed time

is $X_{ik} = min(T_{ik}, C_{ik})$, with the failure indicator $\delta_{ik} = I(X_{ik} = T_{ik})$, where $I(\cdot)$ is the indicator function. Within each cluster, we assume exchangeability among individuals. So the hazard function of T_{ik} given Z_{ik} for the kth individual from the ith cluster is assumed to take the form

$$\lambda_k(t|Z) = \lambda_0(t)exp\{\beta^T Z_{ik}\}, \quad t \geq 0, \quad k = 1, ..., K, \tag{4.1}$$

where β and $\lambda_0(t)$ represent respectively biomarker coefficient parameter and the baseline hazard function. This hazard function essentially drives from the model proposed by Wei et al. (1989), which allows different baseline hazards for different units within each cluster as

$$\lambda_k(t|Z) = \lambda_{0k}exp\{\beta Z_{ik}\}, \quad t \geq 0, \quad k = 1, ..., K. \tag{4.2}$$

In model (4.2) each marginal model has its own regression parameters, whereas in model (4.1) a common set of regression parameters across all K marginal models. A main feature of (4.1) is that the covariate effects on the failures in all marginal models are common and are jointly evaluated. In this article we use (4.1) as hazard model assumption.

The pseudo-partial likelihood proposed by Lee et al. (1992) and Wei et al. (1989)

is

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{exp\{\beta' Z_{ik}(X_{ik})\}}{\sum_{j=1}^{n} \sum_{l=1}^{K} Y_{jl}(X_{ik}) exp\{\beta' Z_{jl}(X_{ik})\}} \right\}^{\delta_{ik}}.$$

under model (4.1) and

$$L(\beta) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{exp\{\beta' Z_{ik}(X_{ik})\}}{\sum_{j=1}^{n} Y_{jk}(X_{ik}) exp\{\beta' Z_{jk}(X_{ik})\}} \right\}^{\delta_{ik}}.$$

under model (4.2) (Note the difference in the denominator between the two functions above). The corresponding score functions $\partial log L(\beta)/\partial\beta$ are

$$U(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{\bar{S}^{(1)}(\beta, X_{ik})}{\bar{S}^{(0)}(\beta, X_{ik})} \right\}.$$

and

$$U(\beta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \left\{ Z_{ik}(X_{ik}) - \frac{S_k^{(1)}(\beta, X_{ik})}{S_k^{(0)}(\beta, X_{ik})} \right\}.$$

where $S_k^{(0)}(\beta, t) = \sum_{j=1}^{n} Y_{jk}(t) e^{\beta' Z_{jk}(t)}$, $S_k^{(1)}(\beta, t) = \sum_{j=1}^{n} Y_{jk}(t) Z_{jk}(t) e^{\beta' Z_{jk}(t)}$ for $k = 1, ..., K$, and $\bar{S}^{(r)}(\beta, t) = \sum_{k=1}^{K} S_k^{(r)}(\beta, t)$ for $r = 0, 1$. In both cases, we obtain the unique estimator of β by solving $U(\beta) = 0$. A sandwich type of variance-covariance estimator is derived from the Taylor expansion of the asymptotic normality of $U(\beta)$ (Wei et al., 1989).

4.2.2 Definition and asymptotic properties of ROC function in independent survival data

Song and Zhou (2008) proposed the estimation of cumulative and incident ROC curves adjusting covariate effect and showed the advantage over Heagerty-Lumley-Pepe and Heagerty-Zheng's approaches in terms of efficiency of estimators using simulation studies. Denoting the survival function for the ith subject with biomarker value X_i at time t as $S(t|X_i) = exp\{-\Lambda_0(t)exp(\beta X_i)\}$ in the proportional hazard model setting, the Song-Zhou ROC function estimations can be expressed as:

$$FPR(x;t) = \frac{\int_x^{\infty} S(t|u)dP(X \leq u)}{\int_{-\infty}^{\infty} S(t|u)dP(X \leq u)} = \frac{\sum_{i=1}^{n} \hat{S}\{t|X_i\}I(X_i \geq x)}{\sum_{i=1}^{n} \hat{S}\{t|X_i\}}, \tag{4.3}$$

$$TPR_C(x;t) = \frac{\int_x^{\infty} \{1 - S(t|u)\}dP(X \leq u)}{\int_{-\infty}^{\infty} \{1 - S(t|u)\}dP(X \leq u)} = \frac{\sum_{i=1}^{n} 1 - \hat{S}\{t|X_i\}I(X_i \geq x)}{\sum_{i=1}^{n} \left[1 - \hat{S}\{t|X_i\}\right]}, \tag{4.4}$$

$$TPR_I(x;t) = \frac{\int_x^{\infty} f(t|u)dP(X \leq u)}{\int_{-\infty}^{\infty} f(t|u)dP(X \leq u)} = \frac{\sum_{i=1}^{n} exp\left\{\hat{\beta}X_i\right\} \hat{S}\{t|X_i\}I(X_i \geq x)}{\sum_{i=1}^{n} exp\left\{\hat{\beta}X_i\right\} \hat{S}\{t|X_i\}}, \tag{4.5}$$

Song and Zhou (2008) proved the asymptotic property of the cumulative and incident ROC functions and estimated the standard error using the bootstrap method. Li and Ning (2015) developed the R package to calculate the ROC functions and corresponding standard error estimators, and we use their package in this article to calculate the

standard error upon resampled data set.

4.2.3 Within-cluster resampling estimation for clustered survival data

Following Hoffman et al. (2001), we randomly sample, with replacement, one individual from each of the n clusters. The bth resampled dataset denoted by $\{X_i^b, \delta_i^b, \mathbf{Z}_i^b(t); i = 1, ..., k, t \in [0, \tau]\}$, consists of n independent observations, which can be analyzed using the Cox proportional hazards model for independent failure time data. For $b = 1, 2, ..., B$, where B is a large fixed number, let $Y_i^b(t) = I(X_j^b \geq t)$ be the survival indicator on whether ith subject in the resampled data set survives at the time t, we introduce the following notation:

$$\begin{aligned}
&\mathbf{S}_b^{(k)}(\beta, t) = n^{-1} \sum_{i=1}^{n} Y_i^b(t) \mathbf{Z}_i^b(t)^{\otimes k} exp\{\beta' \mathbf{Z}_i^b(t)\}, \\
&\mathbf{s}^{(k)}(\beta, t) = E\left\{\mathbf{S}_b^{(k)}(\beta, t)\right\}, \\
&\mathbf{e}(\beta, t) = \frac{\mathbf{s}^{(1)}(\beta, t)}{\mathbf{s}^{(0)}(\beta, t)}, \\
&\mathbf{V}_b(\beta, t) = \frac{\mathbf{S}_b^{(2)}(\beta, t)}{\mathbf{S}_b^{(0)}(\beta, t)} - \left\{\frac{\mathbf{S}_b^{(1)}(\beta, t)}{\mathbf{S}_b^{(0)}(\beta, t)}\right\}^{\otimes 2},
\end{aligned}$$

where $\mathbf{a}^{\otimes k} = 1, \mathbf{a}, \mathbf{a}\mathbf{a}'$ for $k = 0, 1, 2$.

For the bth resampled data, the partial likelihood function is

$$L_b(\beta) = \prod_{i=1}^{n} \left[\frac{exp\{\beta' \mathbf{Z}_i^b(Y_i^b)\}}{\mathbf{S}_b^{(0)}(\beta, Y_i^b)}\right]^{\delta_i^b}, \tag{4.6}$$

and accordingly, the score function is

$$U_b(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \mathbf{Z}_i^b(t) - \frac{\mathbf{S}_b^{(1)}(\beta, t)}{\mathbf{S}_b^{(0)}(\beta, t)} \right\} dN_i^b(t). \tag{4.7}$$

Solving $U_b(\beta) = 0$, we obtain a consistent estimator for β, denoted as $\hat{\beta}_b$. The baseline cumulative hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ can be estimated by the Breslow-Aalan estimator, which for the bth resampled data set is

$$\hat{\Lambda}_0^b(t, \hat{\beta}_b) = \sum_{i=1}^{k} \int_0^t \frac{dN_i^b(u)}{\sum_{j=1}^{k} Y_j^b(u) exp\{\hat{\beta}_b' \mathbf{Z}_i^b(u)\}}.$$

After repeating this procedure B times, the WCR estimator for β, denoted as $\bar{\beta}$, is constructed as the average of the B resample-based estimators,

$$\bar{\beta} = \frac{1}{B} \sum_{b=1}^{B} \hat{\beta}_b, \tag{4.8}$$

and similarly, the WCR estimator for $\Lambda_0(t)$, denoted as $\bar{\Lambda}_0(t, \hat{\beta})$, is

$$\bar{\Lambda}_0(t, \hat{\beta}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\Lambda}_0^b(t, \hat{\beta}_b), \tag{4.9}$$

where $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_B)$.

Under certain regularity conditions (Anderson and Gill (1982), Fleming and Harrington (1990)), for each resampled dataset, $\hat{\beta}_b$ is consistent and asymptotically normal. To prove the asymptotic normality of $\hat{\beta}_b$, the central limit theorem (CLT) cannot

be directly applied because $\bar{\beta}$ is the average of B identically distributed but dependent estimators. Following Hoffman et al. (2001), we can denote $\bar{\beta}$ as the average of m independent cluster-specific terms so that the multivariate CLT can be applied. Here we cite the asymptotic normality property of WCR estimator (Cong et al., 2007).

1. Under regularity conditions, as $n \to \infty$, $\sqrt{n}(\bar{\beta} - \beta_0) \to N_p(0, \Sigma)$ in distribution, where Σ is a finite and positive definite matrix.

2. Under regularity conditions, $\hat{\mathbf{\Sigma}}$ is consistent. As B increases, the covariance matrix of $\bar{\beta}$ converges to Σ. A consistent estimator for Σ is given as

$$\hat{\Sigma} = \frac{n}{B}\left\{\sum_{b=1}^{B}\hat{\Sigma_b} - (B-1)\hat{\Omega}\right\}, \tag{4.10}$$

where $\hat{\Sigma}_b$ is the estimated variance-covariance matrix of $\hat{\beta}_b$ given by

$$\hat{\Sigma}_b = \left\{\sum_{i=1}^{n}\int_0^{\tau}\mathbf{V}_b(\hat{\beta}_b, t)dN_i^b(t)\right\}^{-1},$$

and $\hat{\Omega}$ is the estimated variance-covariance matrix among the B resample-based estimators $\hat{\beta}_b$,

$$\hat{\Omega} = (B-1)^{-1}\sum_{b=1}^{B}(\hat{\beta}_b - \bar{\beta})(\hat{\beta}_b - \bar{\beta})'.$$

3. Let $W(t) = \sqrt{n}\bar{\Lambda}_0(t, \hat{\beta}) - \Lambda_0(t), t \in [0, \tau]$, and let $\mathcal{W}(t)$ be a zero mean Gaussian process with a finite covariance function. The random process $W(t)$ converges weakly

to $\mathcal{W}(t)$ for $t \in [0, \tau]$.

An advantage of the WCR method is that the estimation can be obtained by maximizing the standard partial likelihood function for independent data without specifying any correlation structure. After we obtain the consistent estimator of β and $\lambda_0(t)$ by simply averaging over the estimators from the resampled data, we can also obtain the variance-covariance matrix of $\bar{\beta}$ in a straightforward way as following two sections show.

4.2.4 Definition and asymptotic properties of ROC function in independent survival data

For survival data there are several potential extensions of cross-sectional sensitivity and specificity. Rather than a simple binary outcome, $Y_i = 1$ a survival time can be viewed as a time-varying binary outcome by focusing on the counting process representation $N_i(t) = 1(T_i \leq t)$. Accuracy extensions are classified according to whether the "cases" used to define time-dependent sensitivity are incident cases where $T = t$, or equivalently $dN_i(t) = 1$, is used to define cases for time t, or cumulative cases where $T < 1$ or $N_i^*(t) = 1$ is used. We also consider whether "controls" are static, defined as subjects with $T_i > t^*$ for a fixed value of t^*, or whether controls are dynamic and defined for time t as those subjects with $T_i > t$. Following Heagerty and Zheng (2005) we use the superscripts $\mathbb{C}$ and to denote different definitions of sensitivity, and use the superscripts and to denote different definitions of specificity. In this article we focus on a continuous variable marker X, that is used as a predictor of probability

of failure. When our interest is in the accuracy of a hazard regression model in life science data survival analysis, we use X as the biomarker.

For a baseline marker value, Heagerty et al. (2000) proposed versions of time-dependent sensitivity and specificity using the definitions as

$$sensitivity^{\mathbb{C}}(c,t) : P(X_i > c|T_i \leq t) = P(X_i > c|N_i(t) = 1)$$

$$specificity^{\mathbb{D}}(c,t) : P(X_i \leq c|T_i > t) = P(X_i \leq c|N_i(t) = 0).$$

Using this approach, at any fixed time t the entire population is classified as either a case or a control on the basis of vital status at time t. Also, each individual plays the role of a control for times $t < T$, but then contributes as a case for later times, $t \geq T_i$. ROC curves are defined as $ROC_t^{\mathbb{C}/\mathbb{D}}(p) = TP_t^{\mathbb{C}}\left\{[FP_t^{\mathbb{D}}]^{-1}(p)\right\}$ where $TP_t^{\mathbb{C}}(c) = P(X_i > c \mid N_i^*(t) = 1)$, $FP_t^{\mathbb{D}}(c) = P(X_i > c \mid N_i^*(t) = 0)$ and $[FP_t^{\mathbb{D}}]^{-1}(p) = inf_c c : FP_t^{\mathbb{D}}(c) \leq p$. In the absence of censoring, $ROC_t^{\mathbb{C}/\mathbb{D}}(p)$ can be estimated using the empirical distribution of the marker separately among cases and controls. With censored survival times Heagerty et al. (2000) developed a non-parametric estimator based on either Kaplan-Miere estimator Kaplan and Meier (1958) or the nearest-neighbor bivariate application estimator of Akritas (1994). On the other hand, Etzioni et al. (1999) and Slate and Turnbull (2000) proposed another set of time-dependent sensitivity and specificity as

$$sensitivity^{\mathbb{I}}(c,t) : P(X_i > c \mid T_i \leq t) = P(X_i > c \mid dN_i^*(t) = 1)$$

$$specificity^{\bar{\mathbb{D}}}(c,t) : P(X_i \leq c \mid T_i > t) = P(X_i \leq c \mid N_i^*(t) = 0).$$

where $dN_i^*(t) = N_i^*(t) - N_i^*(t-)$. Using this definition, each subject does not change disease status and is treated as either a case or a control. Cases are stratified according to the time at which the event occurs (incident) and controls are defined as those subjects who are event free through a fixed follow-up period, $(0, t^*)$ (static). These definitions facilitate the use of standard regression approaches for characterizing sensitivity and specificity. Heagerty and Zheng (2005) proposed the *incident/dynamic* definitions of sensitivity and specificity adopting Heagerty et al. (2000) and Etzioni et al. (1999) as

$$sensitivity^{\mathbb{I}}(c,t) : P(X_i > c \mid T_i \leq t) = P(X_i > c \mid dN_i^*(t) = 1)$$

$$specificity^{\mathbb{D}}(c,t) : P(X_i \leq c \mid T_i > t) = P(X_i \leq c \mid N_i^*(t) = 0).$$

Song and Zhou (2008) proposed the estimation of cumulative and incident ROC curves adjusting covariate effect and showed the advantage over Heagerty-Lumley-Pepe and Heagerty-Zheng's approaches in terms of efficiency of estimators using simulation studies. The Song-Zhou ROC function estimations are

$$FPR(x;t) = \frac{\int_y^{\infty} S(t|u)dP(Y \leq u)}{\int_{-\infty}^{\infty} S(t|u)dP(Y \leq u)} = \frac{\sum_{i=1}^{n} \hat{S}\{t|X_i\}I(X_i \geq x)}{\sum_{i=1}^{n} \hat{S}\{t|X_i\}}, \tag{4.11}$$

$$TPR_C(x;t) = \frac{\int_y^{\infty} \{1 - S(t|u)\}dP(Y \leq u)}{\int_{-\infty}^{\infty} \{1 - S(t|u)\}dP(Y \leq u)} = \frac{\sum_{i=1}^{n} 1 - \hat{S}\{t|X_i\}I(X_i \geq x)}{\sum_{i=1}^{n} \left[1 - \hat{S}\{t|X_i\}\right]}, \tag{4.12}$$

$$TPR_I(x;t) = \frac{\int_y^{\infty} f(t|u)dP(Y \leq u)}{\int_{-\infty}^{\infty} f(t|u)dP(Y \leq u)} = \frac{\sum_{i=1}^{n} exp\left\{\hat{\beta}X_i\right\} \hat{S}\{t|X_i\} I(X_i \geq x)}{\sum_{i=1}^{n} exp\left\{\hat{\beta}X_i\right\} \hat{S}\{t|X_i\}}, \quad (4.13)$$

Song and Zhou (2008) proved the asymptotic property of the cumulative and incident ROC functions and estimated the standard error using the bootstrap method. Li and Ning (2015) developed the R package to calculate the ROC functions and corresponding standard error estimators, and we use their package in this article to calculate the standard error upon each resampled data set which is executed in the following section.

4.2.5 Estimate the ROC functions using within cluster resampling method

We apply the within cluster resampling methods for cluster ROC data. Denote $\hat{\alpha}_b$ ($b = 1, ..., B$) as the estimator of parameter α in the bth ($b = 1, ..., B$) resampled data set among the B resampled data sets, the point estimator of α is the average of $\hat{\alpha}$s as

$$\bar{\hat{\alpha}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\alpha}_b,$$

and the variance of $\hat{\alpha}$ is estimated by

$$\widehat{var}(\bar{\hat{\alpha}}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{var}(\hat{\alpha}_b) - \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\alpha}_b - \bar{\hat{\alpha}}\right)^2.$$

Note that $B^{-1} \sum_{b=1}^{B} \widehat{var}(\hat{\alpha}_b)$, which is the consistent estimator of $var(\alpha_b)$, can be

written in the sum of two parts, with one being the average of conditional expectation of variance and the other one being the average of conditional variance of expectation on all the resampled data sets as

$$var(\hat{\alpha}_b) = var\left\{E(\hat{\alpha}_b|data)\right\} + E\left\{var(\hat{\alpha}_b|data)\right\}.$$

In the equation above, the left hand side of the equation can be consistently estimated by $B^{-1}\sum_{b=1}^{B}\widehat{var}(\hat{\alpha}_b)$ as B is large. The first term on the right side hand of the equation is the WCR estimated variance, which is the conditional variance of the expectation of averaging over the resampled variance on all the resampled data sets, because $E(\hat{\alpha}_b\big|data) = \bar{\hat{\alpha}}$. On the other hand, the second term on right side hand of the equation is the average of conditional expectation of variance of the B estimators on all the resampled data sets. We denote $E\left\{var(\hat{\alpha}_b|data)\right\}$ as S_α^2. Therefore, the WCR estimated variance of α equals to $B^{-1}\sum_{b=1}^{B}\widehat{var}(\hat{\alpha}_b)$ subtract S_α^2.

Let α denotes any interested parameter and $\mathcal{P}$ denotes a valid procedure to calculate the parameter α based on $\mathbf{X}$, the following diagram illustrates the steps of our proposed WCR procedure for ROC function estimation.

	$\underline{\mathbf{X}}$				
	$\downarrow$				
1	$\mathbf{X}_{(1)}$	$\rightarrow$	$\mathcal{P}$	$\rightarrow$	$\hat{\alpha}_1, \hat{\sigma}_1^2$
2	$\mathbf{X}_{(2)}$	$\rightarrow$	$\mathcal{P}$	$\rightarrow$	$\hat{\alpha}_2, \hat{\sigma}_2^2$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
B	$\mathbf{X}_{(B)}$	$\rightarrow$	$\mathcal{P}$	$\rightarrow$	$\hat{\alpha}_B, \hat{\sigma}_B^2$
					$\Downarrow$
					$\left(\bar{\hat{\alpha}}, \bar{\hat{\sigma}}^2, S_{\hat{\alpha}}^2\right)$

Note the estimators $\hat{S}(t), \widehat{FPR}, \widehat{TPR}_C, \widehat{TPR}_I, \widehat{ROC}_C$ and $\widehat{ROC}_I$ can be viewed as α in the WCR procedure above. Therefore the point estimators and the corresponding variance estimators of all the ROC parameters can be obtained using the WCR procedure.

4.3 Simulation Study

Extensive simulation studies are conducted to assess the finite sample behavior of the inference procedures proposed in Section 4.2. We consider the marginal model with only one biomarker as covariate in two scenarios of constant cluster size and varying

cluster size respectively.

Clustered failure times are generated using the method proposed by Rader et al. (2014). In this article we perform the ROC function estimation on non-informative and informative cluster size respectively. For non-informative cluster size scenario, we consider the non-informative cluster size as $n_i = k$ for all n clusters. Following steps in Rader et al. (2014), we specify the $k \times k$ correlation coefficients matrix for each cluster in a structure where all the off-diagonal entries are ρ and diagonal entries 1. Then we create a $kn \times kn$ block diagonal matrix, denoted as $\boldsymbol{\Sigma}$ with diagonal entries $\boldsymbol{V}$ and off-diagonal entries all 0. For each $\mathbf{V}_i$, we do the orthogonal decomposition. We denote the eigenvalue that corresponds to $\boldsymbol{V}_i$ as $\Lambda_i = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ and corresponding eigenvectors as $e_{i1}, ..., e_{ik}$, we can decompose $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T$, where $\boldsymbol{\Lambda} = diag(\Lambda_1, ..., \Lambda_n)$ and $\boldsymbol{E}$ is a matrix consist of eigenvectors as $\boldsymbol{E} = (e_{11}, ..., e_{1k}, ..., e_{n1}, ..., e_{nk})$. First, we get the "square root" of $\boldsymbol{\Sigma}$ as $\boldsymbol{E}diag(\lambda_1^{1/2}, \lambda_2^{1/2}, ..., \lambda_{kn}^{1/2})\boldsymbol{E}'$, denoted as $\boldsymbol{R}$. Second, we simulate kn i.i.d. random numbers from standard normal distributed U_{ij} and calculated $\boldsymbol{Y} = \boldsymbol{R}\boldsymbol{U}\boldsymbol{R}^T$ so that we have $\boldsymbol{Y} \sim N_{kn}(\mathbf{0}, \boldsymbol{\Sigma})$. Finally, the baseline survival time for each unit in each cluster is derived from the inverse distribution function of $\Phi(Y_{ij})$, where Φ is the cumulative density function of the standard normal distribution.

To count the biomarker Z's effect into the survival time simulation, first we simulate n i.i.d. standard normal distributed random numbers. Then we expand the size-k cluster by replicating each number k times, for each set identical k numbers we plus normal frailty denoted as $\epsilon_i, i = 1, ..., k$ where $\epsilon \sim N(0, \sigma^2)$. We set $\sigma^2 = 0.225$

in the simulation study. Therefore, the data have correlation coefficient between clusters is 0, whereas within cluster the correlation is close to 1. Here, we mimic the scenario where within cluster the observations that share same environment should have highly correlated biomarker values but observations between clusters should have highly uncorrelated biomarker values. With the assumption that baseline hazard is constant at 0.1, we drive the failure time as $T_{ij} = -10 * log(1 - \Phi(Y_{ij}))exp\{-\beta Z_{ij}\}$. The censoring times, C_{ij}, are generated from the uniform distribution, $Unif(0, c)$, where the value of c can be used to selected to achieve desired censoring rates. We let $c = 50$ corresponding to 10% censoring rate. Final observed survival time is obtained by $X_{ij} = min(T_{ij}, C_{ij})$. We took the number of clusters $m = 200$ or 400. For each setup, we simulated 1000 datasets and analyze each dataset using the WCR method with resampling size, B, following the algorithm proposed by Follmann et al. (2003).

We evaluate the finite-sample performance of the proposed estimator on simulated data and compare the performance to the naive approach which treat the clustered survival times as independent observations. The simulation study results of time-dependent ROC curves for the clustered survival data are summarized respectively in Table 4.1 and Table 4.2. For each data set generated, we obtained the point estimators of cumulative and incident ROC functions evaluated at 0.1, 0.3, 0.3, 0.5 and 0.9 TPR values and report the difference with true estimator which has true coefficient as $\beta = 1$ as, as well as the corresponding standard error using WCR method. We also calculated the sample standard deviation (SD) over the 1000 simulations, the mean standard error (SE) and the 95% confidence interval coverage percentage (CP) for each estimated ROC function. In most of the simulations, the B is between 1600 and

2400. To compare our proposed approach with naive approach, we also reported all the results estimated using naive approach. We can see both approaches yielded the negligible bias. Also, the SE estimations from both WCR and naive approaches are close to empirical SD, but the SE estimated using WCR approach is more closer to the empirical SD than that from naive approach, so as the CP. As shown in Tables 1 and 2, when the cluster size is informative, the point estimates of the ROC functions using the WCR method are approximately unbiased and the 95% confidence interval coverage rates are close to the nominal value, whereas the MM estimates are substantially biased. On the other hand, when the cluster size is non-informative, all point estimates are approximately unbiased and the coverage rates of all three methods are reasonably close to the nominal level. In both tables, the variation of the parameter estimates decreases when the number of clusters increases. The sample SDs are close to the mean SEs for the WCR method over the $(0, 1)$ domain, which suggests that the WCR method provide good estimates for the variability of ROC functions. Contrasted to the proposed WCR method, the naive method which ignores the correlation within cluster generally perform worse in terms of similarity between the sample SDs and the mean SEs, as well as corresponding coverage rate.

Table 4.1: Simulation results for ROC estimators evaluated at $t = 1$ in constant cluster size scenario. Bias is the empirical bias ($\times$ 1000); SD is the empirical standard deviation ($\times$ 1000); SE is the averaged bootstrapping-estimated standard errors ($\times$ 1000); ECR is the exponent of empirical convergence rate.

			Constant Cluster Size							
			Cumulative ROC				Incident ROC			
	Cut-off	Method	Bias	SD	SE	CP	Bias	SD	SE	CP
n=200										
	0.1	WCR	-1.26	40.9	40.6	97.3	-0.33	21.0	18.6	97.0
		Naive	-1.28	41.1	30.9	89.6	-0.37	21.8	15.6	91.3
	0.3	WCR	-0.86	30.2	29.5	96.1	-0.42	24.6	23.4	96.1
		Naive	-0.95	30.3	22.6	90.3	-0.49	24.6	18.8	92.4
	0.5	WCR	-0.12	18.9	18.5	97.0	-0.43	19.0	18.8	96.3
		Naive	-0.15	19.0	14.2	90.7	-0.46	19.3	14.9	93.0
	0.7	WCR	-0.30	9.4	9.3	97.0	-0.46	11.1	11.2	95.4
		Naive	-0.31	9.4	7.2	91.0	-0.47	11.0	8.8	91.2
	0.9	WCR	-0.19	2.4	2.2	94.4	-0.19	3.3	3.0	94.0
		Naive	-0.20	2.4	1.7	88.6	-0.18	3.3	2.5	90.6
n=400										
	0.1	WCR	1.09	29.3	28.0	97.2	1.94	15.0	13.8	93.5
		Naive	1.01	30.4	20.5	90.5	1.99	15.4	11.7	90.5
	0.3	WCR	0.74	21.3	20.2	95.7	0.46	18.1	15.6	93.4
		Naive	0.67	22.2	40.6	87.7	0.62	18.0	12.4	89.6
	0.5	WCR	0.12	13.3	12.7	97.2	0.47	14.3	12.8	96.2
		Naive	0.10	14.3	10.9	88.2	0.54	14.3	11.0	91.9
	0.7	WCR	0.14	6.7	6.3	95.3	0.27	8.3	7.6	96.2
		Naive	0.11	6.9	5.4	91.9	0.32	8.4	7.5	92.9
	0.9	WCR	0.04	1.6	1.6	96.7	0.13	2.4	2.2	97.6
		Naive	0.06	1.7	1.1	89.1	0.18	2.6	1.9	91.3

Table 4.2: Simulation results for ROC estimators evaluated at $t = 1$ in varying cluster size scenario. Bias is the empirical bias ($\times$ 1000); SD is the empirical standard deviation ($\times$ 1000); SE is the averaged bootstrapping-estimated standard errors ($\times$ 1000); ECR is the exponent of empirical convergence rate.

			Varying Cluster Size							
			Cumulative ROC				Incident ROC			
	Cut-off	Method	Bias	SD	SE	CP	Bias	SD	SE	CP
n=200										
	0.1	WCR	-0.95	47.8	45.6	95.7	1.19	25.8	23.5	96.5
		Naive	1.13	45.4	36.9	90.3	1.08	25.4	17.7	90.4
	0.3	WCR	-0.74	21.3	20.2	96.1	0.56	18.1	15.6	96.1
		Naive	-0.88	22.2	15.3	87.7	0.54	18.0	12.4	91.7
	0.5	WCR	0.42	13.3	12.7	97.2	0.67	14.3	12.8	96.0
		Naive	0.30	14.3	11.9	92.2	0.74	14.3	11.0	91.9
	0.7	WCR	0.21	6.7	6.3	95.3	0.27	8.3	7.6	95.9
		Naive	0.18	6.9	6.0	91.9	0.32	8.4	7.5	92.9
	0.9	WCR	0.07	1.6	1.6	96.7	0.13	2.4	2.2	96.1
		Naive	0.09	1.7	1.3	90.1	0.18	2.6	2.2	95.3
n=400										
	0.1	WCR	0.63	22.4	22.9	96.2	1.17	15.4	14.3	95.1
		Naive	0.56	22.0	20.4	92.0	1.14	15.1	11.3	89.7
	0.3	WCR	0.56	12.0	11.3	94.6	0.47	17.4	17.0	97.0
		Naive	0.34	12.3	9.4	90.0	0.45	17.1	14.3	91.8
	0.5	WCR	0.21	9.1	9.9	95.8	0.56	13.6	14.0	96.5
		Naive	0.21	9.3	7.1	88.8	0.56	13.6	11.0	90.5
	0.7	WCR	0.19	4.4	4.4	95.2	0.35	6.6	6.9	95.9
		Naive	0.28	5.8	4.0	88.6	0.16	6.1	4.9	90.0
	0.9	WCR	0.11	2.2	2.1	96.8	0.09	1.1	1.1	94.5
		Naive	0.06	5.8	4.0	90.5	0.11	1.3	0.8	89.7

4.4 Data Examples

We provide an illustration with the well-known Sorbinil Retinopathy Trial (Sorbinil Retinopathy Trial Research Group, 1990), which was conducted between August 1983 and June 1985 to evaluate the effectiveness of the aldose reductase inhibitor to slow the development of diabetic retinopathy. In this study, 497 patients aged 18 to 56 years with insulin-dependent diabetes mellitus for 1 to 15 years were randomly assigned to take oral sorbinil or placebo and followed up for a median of 41 months. The endpoint is two-step progression in retinopathy from baseline on the early treatment diabetic retinopathy study (ETDRS) diabetic retinopathy grading scale. The patients were followed for the occurrence of diabetic retinopathy progression in their left and right eyes. The data set contains following variables: id, eye, survival time (duration from enrollment to the onset of diabetic retinopathy progression or administrative censoring), progression indicator, duration of diabetes at randomization, diastolic blood pressure, total glycated hemoglobin at randomization, sorbinil assignment, cholesterol level. Referring to the literature (Klein and Klein, 2002; Singh et al., 1991), we decide to adjust treatment (sorbinil assignment) and demographic variables (duration of diabetes and cholesterol level) as covariates in the marginal model, and choose two lab test variables, diastolic blood pressure (DBP) and total glycated hemoglobin (TGH), as the biomarkers. In this article, we combine the two biomarkers into one score using the linear combination using coefficients estimated from marginal model as $0.03566 \times DBP + 0.24283 \times TGH$. In this study, each patient could potentially experience diabetic retinopathy progression in both eyes and the time-to-progression endpoints on both eyes for a certain patient can be viewed

as clustered survival data. Therefore, our proposed time-dependent ROC function to evaluate the accuracy of discrimination of a biomarker in clustered survival data can be applied to analyze this data. Since there is no biological difference between the left and right eyes, it is natural to assume a common baseline hazard function for the two failure types. As mentioned, the primary objective of our analysis in this article is to assess whether the biomarker can accurately predict the classification of patients of diabetic retinopathy progression at a certain time potential based on time to onset of progression.

To ensure the validity of these estimators, we checked the proportional hazards assumption using the method in Therneau and Grambsch (2000). There was no evidences against the proportional hazards assumptions. To compare the accuracy of the proposed score as the biomarker in distinguishing/predicting the subjects experience diabetic retinopathy progression by/at a given time t and those experience diabetic retinopathy progression after t, we estimated the cumulative and incident ROC curves separately for the two biomarkers using the estimator $\widehat{ROC}_C$ and $\widehat{ROC}_I$ adjusting sorbinil assignment, duration of diabetes and cholesterol level at various time points after enrollment. In Figure 3.1, we plot the estimated the cumulative and incident ROC curves for the proposed score marker at $t = 1, 2$ and 4 years after enrollment to compare the classification/prediction accuracy of the proposed score marker distinguishing subjects. The 95% point-wise confidence intervals estimated based on the robust estimator of standard error were computed by the WCR method using 2000 resampled data sets. The point-wise confidence intervals estimated using naive approach are also presented in the plots for reference. From the plots we can

see the WCR estimated confidence intervals are wider than those estimated using naive approach. Generally, the proposed score that combines diastolic blood pressure and total glycated hemoglobin works well in classifying cases and controls for various time points in that the estimated ROC curves and their confidence intervals are above the diagonal line at all of the 3 time points. The objective of the present analysis is twofold. First we use the proposed combined score as the marker to evaluate its performance to discriminate between subjects who experienced diabetic retinopathy progression by t years versus those who were progression free by t years using cumulative ROC function. This objective can be interpreted as to identify those subjects who are at "high risk" and for whom intervention is warranted. We also look at the predictive ability of this marker to distinguish subjects who experience progression at t years versus those who are progression free by t years using incident ROC function. This objective can be interpreted as to identify (to treat) those subjects who are still progression free, but likely to experience progression in the near future. The former objective is evaluated using the cumulative ROC function that is based on baseline or time-independent marker, while the later is evaluated by the incident ROC function that is based on time-dependent marker. From the results we see the proposed score can be used for both purposes in terms of its discrimination (cumulative ROC) or prediction (incident ROC) ability, and it is not very sensitive to time t.

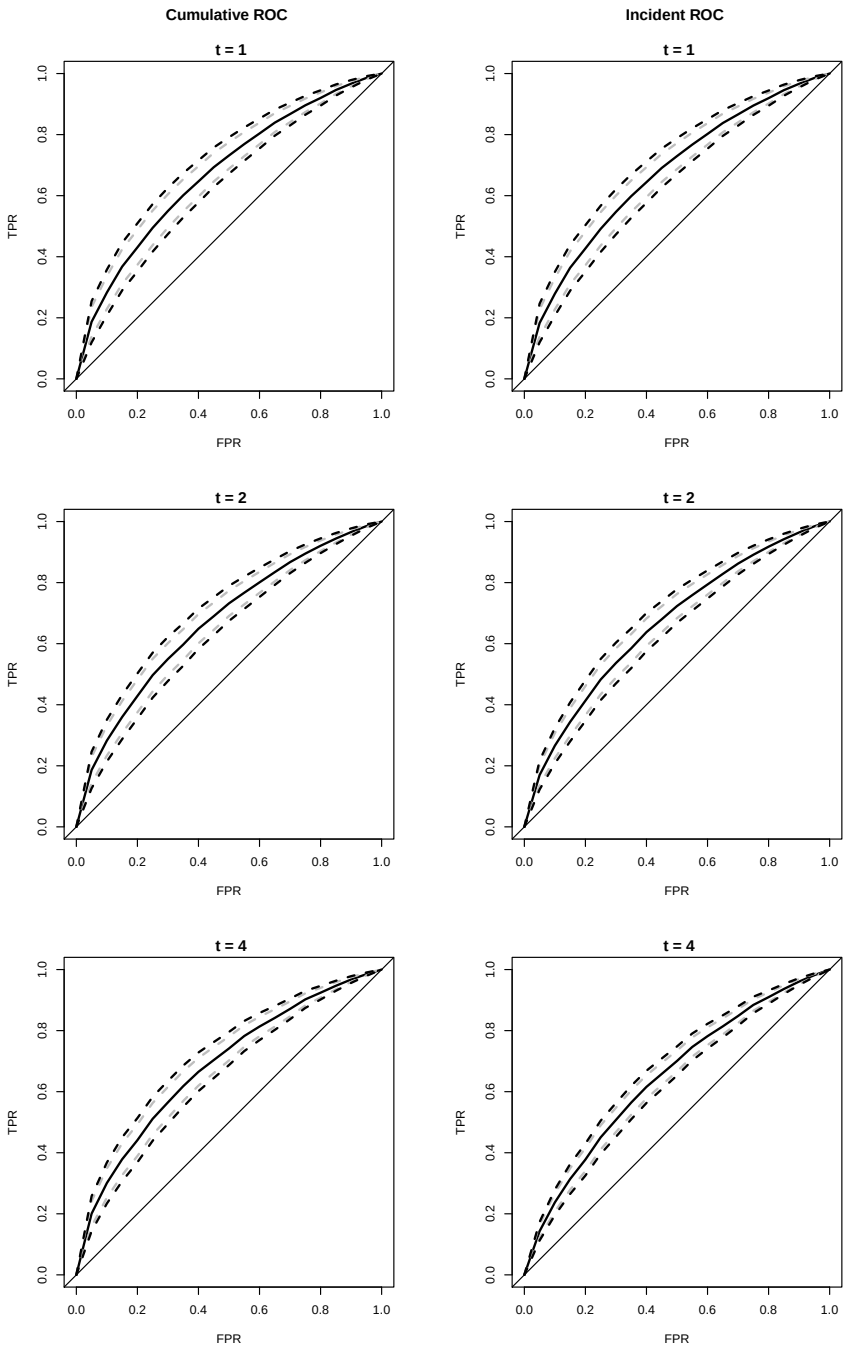

Figure 4.1: Estimated cumulative and incident ROC curves for the proposed biomarker using Sorbinil Retinopathy Trial data. The plots are, from the top, for t = 1, 2 and 4, respectively. Estimated ROC curves is solid lines. Estimated 95% point-wise confidence intervals using WCR and naive approaches are presented as dashed lines by black and gray colors respectively.

4.5 Discussion

The WCR method applied in estimation of ROC curves of clustered survival data based on marginal model estimation has been shown to be unbiased and consistent in the preceding paragraphs. Thanks to the remarkably developed computer science, the simulation based WCR technique can be applied in personal computer without severe burden of time consuming. While there has been methods, such as two-stage bootstrap procedure, implemented in clustered, hierarchical or multilevel data (Cheng et al., 2013; Sherman and Cessie, 1997), the WCR approach possesses the merit of straightforward operation and the type I error control feasibility when determining the resmapling size. The unbiasedness and consistency depends on the correct specification and assumption of proportional hazard model, and the interpretation of our method is limited in population level. As Lin (1994) pointed out, there has been considerable controversy over the unconditional specification of the marginal hazard since β generally needs to be interpreted conditionally on the unobserved frailty. Extension of our method to incorporate estimate beyond marginal model, e.g. frailty model, parameters could constitute a future study. Also, formal comparison of time-depend ROC curves based on area under curve can be pursued, either via parametric or simulation approach.

Chapter 5

Discussion

Variable selection plays an important role in life science research. To a large extent, the validity of scientific inference depends on the correct specification of the model. In a practical data analysis, the analyst has to decide whether a variable should be included in the model, what functional form it should take, and how accurate this covariate can distinguish disease/non-disease status of the subjects. The complexity of the competing risks has greatly complicated the selection process. In this dissertation, by placing a penalty on model complexity, the method fundamentally simplifies the selection process to facilitate simultaneous and automatic variable selection. And after the selection, ROC function serves as the scale-free tool to evaluate the prognostic potential of the selected variables by focusing on the correct classification capability. A noticeable gap in the existing literature is the lack of selection procedures for complex survival data, such as competing risks, interval censoring, correlated time etc. The increasing popularity of the sub-distribution hazards model present an urgent demand to fill this gap. This dissertation addresses this need in a systematic way, by proposing an integrated procedure that shows an example to address these demand. In this chapter, I would review the main methodological contribution and practical impact of this research and highlight the meaningfulness and advantages of the three topics in the following paragraphs.

First, this dissertation has presented the method to automatically and simultaneously select linear and non-linear effects for the sub-distribution hazards model. While the selection of the linear effects helps to identify independent variables that are related to the outcomes, selection of the non-linear effects serves the dual purpose of specifying the underlying non-linear effect and provide more insight of the understanding of the effect. Importantly, the reparametrization by spectral decomposition allows the covariate effects in the linear and non-linear components to retain their own covariance structures, while not restricting the model space by pre-excluding candidate models. Such an approach thus enables researchers to simultaneously perform effect selection by identifying their corresponding functional forms. Additionally, the reparametrization also allows the non-linear effects in the sub-distribution hazards models to be linked in a common structure with same parameter dimension. Practically, this reparametrization through spectral decomposition has made the selection of non-linear effects by group penalty feasible. This reparametrization and the linear/non-linear effect selection is not restricted to the sub-distribution hazards model setting for studying the correlation between the covariate effects and the survival outcomes. Actually, it could be extended to any model settings with linear combination of covariates and survival outcomes to investigate their correlations, which should have wide applicability in clinical investigations.

Secondly, this thesis developed a method to identify the functional forms of independent variables in an additive model. It provides a general nonparametric framework for structural discovery in such a model setting. The decomposition of the B-spline basis clearly partitions the independent variable effects into a parametric (linear)

part and a nonparametric (nonlinear) part. We then present the model in a mixed-effect model formulation. Methodologically, the basis decomposition and mixed model representation serve as a bridge between variable selection and structural discovery. Practically, it clearly depicts the independent variable effects as linear and nonlinear other than lumping them together, thus retaining the model interpretability. The same approach could be similarly extended to other survival models such as cause-specific hazards model.

Thirdly, this thesis shows a way to apply the existing techniques in the complex data settings. For example, the within-cluster-resampling technique is originally developed for the longitudinal data analysis, and we apply the technique in the clustered data setting for ROC estimation to avoid complicated variance estimator derivation. This though can be applied in many other scenarios where the estimation has been well developed for independent data but difficult to extend to correlated setting.

Finally, this dissertation presents a general computational strategy for ROC function with competing risks data. The ROC function estimators based on the estimated CIF has its advantage in clear definition and useful interpretation in complex survival setting. The ROC function estimated from the CIF has its own restraint of sum up to 1, so that the ROC function can be correctly used in the scenario where the classification capability of a certain event has been adjusted for the existing of its competing events. The non-parametric B-spline based estimation of the CIF assures the flexibility of the computing while maintain the parameter interpretability and computation feasibility. Furthermore, the application of the non-parametric variable

selection and ROC function estimation is adaptable to some widely used existing statistical packages with affordable computing burden in intermediate variable size. At the conclusion of this dissertation, I am confident that the proposed procedure will achieve more popularity in application. The development of more sophisticated and easier to use packages for implement of the methods will further strengthen the applicability. The methods here are mainly depicted but shall not be limited in the joint model setting.

I anticipate that further modifications and extensions of the current work will become necessary. Future extensions could include variable selections for time dependent covariate in competing risks modeling. Dealing with the missing data is an important aspect that I did not study in the current dissertation. As well as the large sample behavior of the time-dependent ROC function estimations on interval censored data. Notwithstanding these limitations, I hope that increased application of these procedures will stimulate new thinking for the improvement of the proposed methods.

Chapter 6

Appendix

To derive the consistency of the estimators, besides the conditions specified in Section 3.2, we assume the following regularity conditions.

A. T and C are independent given Z.

B. $P(V \geq L) > 0$ for a constant $L > 0$.

C. $E(Z^T Z) < \infty$.

D. $\Phi(t)$ is bounded and has bounded first and second order function for $t \in (-\infty, +\infty)$.

E. For $z \in \mathcal{Z}$, $F(t|z)$ is an absolutely continuous function for $t \in [0, L]$.

F. The conditional densities

$$f^C(z;t) = -\frac{dTPR^C(z;t)}{dz} = \frac{F(t|z)F'(z)}{\int_{-\infty}^{\infty} \{F(t|u)\} dP(Z \leq u)},$$

$$f^I(z;t) = -\frac{dTPR^I(z;t)}{dz} = \frac{f(t|z)F'(z)}{\int_{-\infty}^{\infty} f(t|u) dP(Z \leq u)},$$

$$f_0(z;t) = -\frac{dFPR(z;t)}{dz} = \frac{[1 - F(t|z)]F'(z)}{\int_{-\infty}^{\infty} [1 - F(t|u)] dP(Z \leq u)},$$

exist.

With some algebra, the ROC function can be estimated as :

$$\widehat{TPR}_C(z;t) = \frac{\sum_{i=1}^{n} \hat{F}(t|Z_i) * I\{Z_i > c\}}{\sum_{i=1}^{n} \hat{F}(t|Z_i)}, \tag{6.1}$$

$$\widehat{TPR}_I(z;t) = \frac{\sum_{i=1}^{n} \hat{F}(t|Z_i) * (1 + exp\{\hat{\phi}(t) + \hat{\beta} Z_i\})^{-1} * I\{Z_i > c\}}{\sum_{i=1}^{n} \hat{F}(t|Z_i) * (1 + exp\{\hat{\phi}(t) + \hat{\beta} Z_i\})^{-1}}, \tag{6.2}$$

$$\widehat{FPR}(z;t) \ = \frac{\sum_{i=1}^{n} (1 - \hat{F}(t|Z_i) * I\{Z_i > c\}}{\sum_{i=1}^{n} (1 - \hat{F}(t|Z_i)}, \tag{6.3}$$

where $\hat{F}(t|Z_i) = \frac{exp\{\hat{\phi}(t)+\hat{\beta}Z_i\}}{1+exp\{\hat{\phi}(t)+\hat{\beta}Z_i\}}$. According to Bakoyannis et al. (2016), we know that $\hat{\phi}(t) \xrightarrow{P} \phi(t)$ and $\hat{\beta} \xrightarrow{P} \beta$. According to the continuous mapping theorem of consistent estimator, if we denote (ϕ, β) as $\boldsymbol{\theta}$ and $(1, Z)$ as $\boldsymbol{X}$ we have $\frac{exp\{\hat{\phi}_t+\hat{\beta}Z\}}{1+exp\{\hat{\phi}_t+\hat{\beta}Z\}} \xrightarrow{P} \frac{exp\{\phi_t+\beta Z\}}{1+exp\{\phi_t+\beta Z\}}$, since function $g(\theta) = \frac{exp\{\boldsymbol{X\theta}\}^{\boldsymbol{T}}}{1+exp\{\boldsymbol{X\theta}^T\}}$ is a real-valued function continuous at $\boldsymbol{X}$. Therefore we prove $\hat{F}(t|Z) \xrightarrow{P} F(t|Z)$.

Following Song and Zhou (2008), we have the same regularity condition of (β, Φ). Since FPR is differentiable as a composite functional of (β, Φ), using the functional

Taylor expansion, we have

$$\begin{aligned}
\widehat{TPR}_C(z;t) - TPR_C(z;t) = & \left[\int_{-\infty}^{\infty} F(t|u)d\Phi(u)\right]^{-1} \left[\int_z^{\infty} \left\{\hat{F}(t|u) - F(t|u)\right\} d\Phi(u)\right. \\
& \left. + \int_z^{\infty} F(t|u)d\{\hat{\Phi}(u) - \Phi(u)\}\right] \\
& - \left[\int_{-\infty}^{\infty} F(t|u)d\Phi(u)\right]^{-2} \int_z^{\infty} F(t|u)d\Phi(u) \\
& \times \left[\int_{\infty}^{\infty} \left\{\hat{F}(t|u) - F(t|u)\right\} d\Phi(u)\right. \\
& \left. + \int_{\infty}^{\infty} F(t|u)d\left\{d\hat{\Phi}(u) - \Phi(u)\right\}\right] \\
& + o_p(1).
\end{aligned}$$

Since $\hat{\Phi}(t) \xrightarrow{P} \Phi(t)$ and $\hat{F}(t) \xrightarrow{P} F(t)$ for given t, we have $\hat{\Phi}(t) - \Phi(t) = o_p(1)$ and $\hat{F}(t) - F(t) = o_p(1)$, all the integrals above are converge to 0 in probability. We have the above $\widehat{TPR}_C(z;t) - TPR_C(z;t) \xrightarrow{P} o_p(1)$. Therefor, we prove $\widehat{TPR}_C(z;t) \xrightarrow{P} TPR_C(z;t)$.

Similarly, the consistency of $\widehat{FPR}$ can be proved.

Since the ROC function is the composite of $TPR_C(z;t)$ and $FPR(z;t)$, the consistency can be proved by continuous mapping theorem of consistent estimator.

BIBLIOGRAPHY

Akritas, M. G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics 22*, 1299–1327.

Anderson, P. and R. Gill (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics 10*, 1100—-1120.

Androulakis, E., C. Koukouvinos, and F. Vonta (2012). Estimation and variable selection via frailty models with penalized likelihood. *Statistics in Medicine 29*, 2453–2468.

Bakoyannis, G., M. Yu, and C. T. Yiannoutsos (2016). Semiparametrically efficient estimation for a general class of cumulative incidence regression models with interval-censored data. *Statitstics in Medicine 10*, 1100—-1120.

Belot, A., M. Abrahamowicz, L. Remontet, and R. Giorgi (2010). Flexible modeling of competing risks in survival analysis. *Statistics in Medicine 29*, 2453–2468.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statitstics in Medicine 2*, 273—-277.

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics 41*(2), 802–837.

Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher (2009). Simulating competing risks data in survival analysis. *Statistics in Medicine 28*, 956–971.

Cai, J. and P. RL (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika 82*, 151–164.

Cai, J. and P. RL (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Anal 3*, 197–213.

Cai, T., M. Pepe, T. Lumley, Y. Zheng, and N. Jenny (2006). The sensitivity and specificity of markers for event times. *Biostatistics 7*, 182–197.

Chen, K., Z. Jin, , and Z. Ying (2002). Semiparametric analysis of transformation models with censored data. *Biometrika 89*, 659—-668.

Chen, Y., K. Chen, and Z. Ying (2010). Analysis of multivariate failure time data using marginal proportional hazards model. *Statistica Sinica 20*, 1025–1041.

Cheng, G., Z. Yu, and J. Huang (2013). The cluster bootstrap consistency in generalized estimating equations. *Journal of Multivariate Analysis 115*, 33–47.

Choi, S. and X. Huang (2014). Maximum likelihood estimation of semiparametric mixture component models for competing risk data. *Biometrics 70*, 588—598.

Cong, X., G. Yin, and Y. Shen (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics 63*, 663–672.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society 34*, 187–220.

Crowder, M. (2001). *Classical competing risks*. CRC Press, London.

Dignam, J., Q. Zhang, and M. Kocherginsky (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research 18*, 2301—-2308.

Ding, J. and J. Wang (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics 64*, 546–556.

Eriksson, F., J. Li, T. Scheike, and Z. M (2015). The proportional odds cumulative incidence model for competing risks. *Biometrics 71*, 687—-695.

Etzioni, R., M. Pepe, G. Longton, C. Hu, and G. Goodman (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making 19*, 242–251.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and R. Li (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics 30*, 74–99.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica 20*, 101–148.

Fan, X., I. Gijbels, and M. King (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist. 25*, 1661–1690.

Feng, S., R. Wolfe, and F. Port (2005). Frailty survival model analysis of the national deceased donor kidney transplant dataset using poisson variance structures. *Journal of the American Statistical Association 100*, 728–735.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities. *Biostatistics 2*, 85–97.

Fine, J. P. and R. J. Gray (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association 94*, 496–509.

Fleming, T. and D. Harrington (1990). *Counting Processes and Survival Analysis*. New York: Wiley.

Follmann, D., M. Proschan, and E. Leifer (2003). Multiple outputation: Inference for complex clustered data by averaging analyses from independent data. *Biometrics 59*, 420–429.

Gaynor, J., E. Feuer, and C. Tan (1993). On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association 88*, 400–409.

Gray, R. (1992). Methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association 87*, 942–951.

Gray, R. J. and Y. Li (2002). Optimal weight functions for marginal proportional hazards analysis of clustered failure time data. *Lifetime Data Anal 8*, 5–12.

Ha, I., M. Lee, S. Oh, J. Jeong, R. Sylvester, and Y. Lee (2014). Variable selection in subdistribustion hazard frailty models with competing risks data. *Statistics in medicine 33*, 4590–4604.

He, Z., W. Tu, S. Wang, H. Fu, and Z. Yu (2014). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics 71*, 178–187.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*, 337–344.

Heagerty, P. J. and Y. Zheng (2005). Survival model predictive accuracy and ROC curves. *Biometrics 61*, 92–105.

Hoffman, E., P. Sen, and C. Weinberg (2001). Within-cluster resampling. *Biometrics 88*, 1121–1134.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag.

Huang, J. and A. J. Rossini (1997). Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association 02*, 960–967.

Huang, J. and J. A. Wellner (1997). Interval censored survival data: A review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics 1*, 123—-169.

Hughes, M. D. (1995). Power considerations for clinical trials using multivariate time-to-event data. *Manuscript*.

Jeong, J. H. and J. P. Fine (2007). Parametric regression on cumulative incidence function. *Brain Imaging Behav 8*, 184—-196.

Jones-Davis, D. M. and N. Buckholtz (2015). The impact of the alzheimer's disease neuroimaging initiative 2: What role do public-private partnerships have in pushing the boundaries of clinical and basic science research on alzheimer's disease? *Alzheimer's & Dementia: The Journal of the Alzheimer's Association 11*(7), 860—-864.

Kalbfleisch, J. and R. Prentice (2002). *The Statistical Analysis of Failure Time Data (Second Edition)*. Wiley: New Jersey.

Kaplan, E. and P. Meier (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assn. 53*(282), 457–481.

Klein, R. and B. E. K. Klein (2002). Blood pressure control and diabetic retinopathy. *Br. J. Ophthalmol. 86*(4), 365–367.

Kuk, D. and R. Varadhan (2013). Model selection in competing risks regression. *Statistics in Medicine 32*, 3077–3088.

Lau, B., S. Cole, and S. Gange (2009). Competing risk regression models for epidemiologic data. *Am J Epidemiol 170*, 244–256.

Lee, E. W., L. J. Wei, and D. Amato (1992). *Cox-type regression analysis for large number of small groups of correlated failure time observations*. Kluwer Academic Publishers.

Li, C. (2016). The fine-gray model under interval censored competing risk data. *Journal of Multivariate Analysis 143*, 327—344.

Li, S. and Y. Ning (2015). Estimation of covariate-specific time-dependent roc curves in the presence of missing biomarkers. *Biometrics 73*, 666–676.

Liang, K. Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Lin, D. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine 13*, 2233–2247.

Miao, Z. (2014). Within-cluster resampling methods for clustered receiver operating characteristic (ROC) data. *PhD thesis, Department of Statistics, George Mason University,Fairfax, VA*.

Mueller, S., M. Weiner, L. Thal, R. Petersen, C. Jack, W. Jagust, J. Trojanowski, A. Toga, and L. Beckett (2017). The alzheimer's disease neuroimaging initiative. *Neuroimaging clinics of North America 15*(4), 869.

Murphy, S., A. Rossini, and v. d. A. Vaart (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association 92*, 968—976.

O'Quigley, J. and R. Xu (2001). Explained variation in proportional hazards regression. in handbook of statistics in clinical oncology. *J. Crowley (ed), . New York: Marcel Dekker*, 397—-409.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal of Scientific and Statistical Computing 9*, 531–542.

O'Sullivan, F. (1993). Nonparametric estimation in the cox model. *Annals of Statistics 21*, 124–145.

Pepe, M., Y. Zheng, Y. Jin, Y. Huang, C. Parikh, and W. Levy (2008). Evaluating the roc performance of markers for future events. *Lifetime Data Analysis 14*, 86—-113.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika 84*(3), 595–608.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics 54*, 124–135.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford.

Pepe, M. S. and J. Cai (1993). Some graphical displays and marginal regression analysis for recurrence failure times and time-dependent covariates. *Journal of the American Statistical Association 88*, 811–820.

Pepe, M. S., W. Leisenring, and C. Rutter (1999). *Evaluating diagnostic tests in public health. In Handbook of Biostatistics, Volume 18, C. R. Rau and P. K. Sen (eds)*. New York : Elsevier Scientific.

Prentice, R., J. Kalbfleisch, A. Peterson, N. Flournoy, V. Farewell, and N. Breslow (1978). The analysis of failure times in the presence of competing risks. *Biometrics 34*, 541—-554.

Prentice, R. L. and L. Hsu (1997). Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika 79*, 495–512.

Rader, K., S. Lipsitz, D. Harrington, and M. Parzen (2014). Simulating clustered survival data with proportional hazards margins.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2015). Package 'fda'.

Robins, J. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. *In proceedings of the Biopharmaceutical Section, American Statistical Association*, 24–33.

Saha, P. and P. J. Heagerty (2010). Time-dependent predictive accuracy in the presence of competing risks. *Biometrics 66*, 999—-1011.

Schemper, M. and R. Henderson (2000). Predictive accuracy and explained variation in cox regression. *Biometrics 56*, 249—255.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Sherman, M. and S. l. Cessie (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *J. R. Statist. Soc. B 26*, 901—-925.

Shi, H., Y. Cheng, and J. J. H. (2013). Constrained parametricmodel for simultaneous inference of two cumulative incidence functions. *Biometrical Journal 55*(1), 82–96.

Singh, R., V. Prakash, P. K. Shukla, S. Gautam, and O. P. Maurya (1991). Glycosylated hemoglobin and diabetic retinopathy. *Ann. Ophthalmol. 23*(8), 308–311.

Skinner, J., J. O. Carvalho, G. G. Potter, A. Thames, E. Zelinski, P. K. Crane, and L. E. Gibbons (2012). The alzheimer's disease assessment scale-cognitive-plus (adas-cog-plus): an expansion of the adas-cog to improve responsiveness in mci. *Brain Imaging Behav 6*(4), 10.

Slate, E. H. and B. W. Turnbull (2000). Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine 19*, 617–637.

Song, X. and X. Zhou (2008). A semiparametric approach for the covariate specific ROC curve with survival outcome. *Statistica Sinica 18*, 947–965.

Sorbinil Retinopathy Trial Research Group (1990). A randomized trial of sorbinil, an aldose reductase inhibitor, in diabetic retinopathy. *Arch Ophthalmol 180*(9), 1234–1244.

Speed, T. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science 6*, 42–44.

Steinberg, J., A. Sadaniantz, J. Kron, A. Krahn, D. Denny, J. Daubert, W. Campbell, E. Havranek, K. Murray, B. Olshansky, G. O'Neill, M. Sami, S. Schmidt, R. Storm, M. Zabalgoitia, J. Miller, M. Chandler, E. Nasco, and H. Greene (2004). Analysis of cause-specific mortality in the Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) study. *Circulation 109*(16), 1973–1980.

The AFFIRM Investigators (2002). A comparison of rate control and rhythm control in patients with atrial fibrillation. *N Engl J Med 34*, 1825—1833.

The AFFIRM Investigators (2004). Relationships Between Sinus Rhythm, Treatment and Survival in the Atrial Fibrillation Follow-Up Investigation of Rhythm Management (AFFIRM) Study. *Circulation 109*(12), 1509—1513.

Therneau, T. M. and P. M. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society 58*, 267–288.

Tosteson, A. N. A. and C. B. Begg (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making 8*(3), 204–215.

Wand, M. P. and J. T. Ormerod (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics 50*, 179–198.

Wang, H., R. Li, and C. L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 553–568.

Wei, L. J., D. Y. Lin, and L. Weissfeld (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association 84*, 1065–1073.

Weiner, M., P. Aisen, C. Jack, W. Jagust, J. Trojanowski, L. Shaw, A. Saykin, J. Morris, and N. Cairns (2010). The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association 6*(3), 202—-211.

Yan, J. and J. Huang (2012). Model selection for cox models with time-varying coefficients. *Biometrics 16*, 419–428.

Yang, Y. and Z. Ying (2001). Marginal proportional hazards models for multiple event-time data. *Biometrika 88*, 581–586.

Yoo, W., R. Mayberry, S. Bae, K. Singh, Q. He, and J. Lillard (2014). A study of effects of multicollinearity in the multivariable analysis. *Int J Appl Sci Technol 4*, 9–19.

Yu, Z. and X. Lin (2008). Nonparametric regression using local kernel estimating equations for correlated failure time data. *Biometrika 95*, 123–137.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society 68*, 49–67.

Zhang, H., G. Cheng, and L. Y (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association 106*, 1099–1112.

Zhang, H. and W. Lu (2007). Adaptive Lasso for Cox's proportinal hazards model. *Biometrika 94*, 691–703.

Zhang, Y., L. Hua, and J. Huang (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics 37*, 338—-354.

Zheng, Y., T. Cai, Y. Jin, and Z. Feng (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics 68*, 388–396.

Zheng, Y., T. Cai, J. Stanford, and Z. Feng (2010). Semiparametric models of time-dependent predictive values of prognostic biomarkers. *Biometrics 66*, 50—-60.

Zheng, Y. and P. Heagerty (2007). Prospective accuracy for longitudinal markers. *Biometrics 63*, 332—341.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association 101*, 1418–1429.

Zuo, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society 67*, 301—-320.

Zweig, M. and G. Campbell (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry 88*, 581–586.

CURRICULUM VITAE

Xiaowei Ren

EDUCATION

- Ph.D. in Biostatistics, Indiana University, Indianapolis, IN, 2017 (minor in Epidemiology)

WORKING EXPERIENCE

- Statistician Intern, Allergan Inc., California, U.S.A. May. 2016 - Aug. 2016
- Research Assistant, Indiana University, Indiana, U.S.A. May. 2013 - May. 2017

SELECT PUBLICATIONS

- Wang J, Morale SE, Ren X; Birch EE. (2016) Longitudinal Development of Refractive Error in Children With Accommodative Esotropia: Onset, Amblyopia, and Anisometropia. *Investigative Ophthalmology and Visual Science.* 57: 2203–2212.
- Liangpunsakul S, Puri P, Shah VH, Kamath P, Sanyal A, Urban T, Ren X, Katz B, Radaeva S, Cudd TA, Chalasani N, Crabb DW. Effects of Age, Sex, Body Weight, and Quantity of Alcohol Consumption on Occurrence and Severity of Alcoholic Hepatitis. Translational Research and Evolving Alcoholic Hepatitis Treatment Consortium. *Clin Gastroenterol Hepatol.* 2016 Jun 15.

- Birch SM, Lenox MW, Kornegay JN, Shen L, Ai H, Ren X, Goodlett CR, Cudd TA, Washburn SE. (2015) Computed tomography assessment of peripubertal craniofacial morphology in a sheep model of binge alcohol drinking in the first trimester. *Alcohol.* 49: 675–689.
- Gough G, Heathers L, Puckett D, Westerhold C, Ren X, Yu Z, Crabb DW, Liangpunsakul S. (2015) The Utility of Commonly Used Laboratory Tests to Screen for Excessive Alcohol Use in Clinical Practice. *Alcohol Clin Exp Res.* 39: 1493-1500.
- Wang J, Ren X, Shen L, Yanni SE, Leffler JN, Birch EE. (2013) Development of refractive error in individual children with regressed retinopathy of prematurity. *Investigative Ophthalmology and Visual Science.* 54: 6018–6024
- Shen L, Ai H, Liang Y, Ren X, Anthony CB, Goodlett CR, Ward R, Zhou FC. (2013) Effect of prenatal alcohol exposure on bony craniofacial development: a mouse MicroCT study. *Alcohol.* 47: 405–415.